EX LIBRIS
UNIVERSITATIS
ALBERTENSIS

The Bruce Peel
Special Collections
Library

Digitized by the Internet Archive
in 2025 with funding from
University of Alberta Library

https://archive.org/details/0162017486315

University of Alberta

Library Release Form

Name of Author: Michael Paul Kowalski

Title of Thesis: Non Parametric Sequential Analysis in Clinical Trials with Censored Data

Degree: Master of Science

Year this Degree Granted: 2003

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Through The Never

All that is, was and will be
Universe much too big to see

Time and space never ending
Disturbing thoughts, questions pending
Limitations of human understanding

Too quick to criticize
Obligation to survive
We hunger to be alive

In the dark, see past our eyes
Pursuit of truth no matter where it lies

Gazing up to the breeze of the heavens
On a quest, meaning, reason
Came to be, how it begun
All alone in the family of the sun
Curiosity teasing everyone
On our home, third stone from the sun

Hetfield, Ulrich, Hammett

University of Alberta

Non Parametric Sequential Analysis in Clinical Trials with Censored Data

by

Michael Paul Kowalski

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

in

Statistics

Department of Mathematical and Statistical Sciences

Edmonton, Alberta

Fall 2003

UNIVERSITY OF ALBERTA

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Non-parametric sequential analysis in clinical trials with censored data** submitted by **Michael Paul Kowalski** in partial fulfillment of the requirements for the degree of Master of Science in Statistics.

Dedication

I dedicate this to my parents. I love you both so much. There is no way I would be the person I am today if it wasn't for you both.

Abstract

There are various testing procedures used in clinical trials with censored data. Some of these procedures are group sequential, analyzing the data at a pre-selected number of equal time intervals. Others are pure sequential, analyzing the data at every event time. The main purpose of sequential analysis is the essential need to terminate a study as soon as possible. For obvious financial and ethical reasons, early stopping is beneficial. In this thesis, non-parametric sequential methods will be evaluated with comparison of the group sequential approach to the fully sequential approach. We wish to compare the results in terms of power and average stopping time.

Acknowledgements

I wish to first thank my supervisors, Dr. Edit Gombay and Dr. Giseon Heo, for your guidance throughout this journey. Your knowledge and inspiration have guided me successfully.

I wish to thank Dr. Byron Schmuland and Dr. Paul Major for taking time out of your schedules to be on my examination committee.

As well, I am extremely grateful to Dr. Schmuland and the Department of Mathematical and Statistical Sciences for your financial support. This would not have been completed otherwise.

I would like to thank the Department of Mathematical and Statistical Sciences and the Faculty of Science, especially Rick Mikalonis and Dr. Ivan Baggs for giving me the opportunity to teach lectures and labs. This experience is invaluable.

I wish to thank everyone in the math office for your kindness and helpful hearts. You are all wonderful people to work with.

To my friends, I thank you for always being here for me whenever I've needed you. You all know who you are, and you will be thanked individually.

To everyone who has ever asked me, Aren't you done yet?, I thank you for reminding me that I wasn't. Your constant nagging turned out to be a blessing.

Last, but most importantly, I wish thank my family. To my parents and my sister, your love and constant encouragement made this happen. I love you all.

To anyone I've missed, I apologize.

Contents

1 Introduction 1

1.1 The Thesis 1

1.2 The Title 2

1.2.1 Non Parametric... 2

1.2.2 ...Sequential Analysis... 3

1.2.3 ...in Clinical Trials with Censored Data. 4

1.3 The Hypothesis Test 5

1.4 A Statistical View of the Clinical Trial 8

2 Literature Review 13

2.1 U-Statistics 14

2.2 Wilcoxon Type Statistics 14

2.2.1 Mann-Whitney U statistics 16

2.2.2 Gehan Test for censored data 18

2.3 The Log-Rank test 22

3 Sequential Designs 26

3.1 A General Approach to Group Sequential Testing 27

3.1.1 Tests With Equal Group Sizes 28

3.1.2 Normal Responses 31

3.1.3 Exponential Responses 33

3.1.4 Weibull Responses 36

3.1.5 Log Rank Test 39

3.2 Nominal Significance Level 41

3.2.1 Error Calculation 42

3.2.2 Pocock 43

3.2.3 O'Brien and Fleming 44

3.3 A Pure Sequential Design 46

4 Simulation and Results **50**

4.1 The Simulation 50

4.1.1 The Algorithm 51

4.2 Case 1: Comparison using U-Statistics 53

4.2.1 Results for Comparison 1 54

4.3 Comparison 2: U-Statistic vs. Sequential Log-Rank 60

4.3.1 The Truncation Point, n 63

4.3.2 Sequential Log Rank Test 65

4.3.3 Calculating the Truncation Point, An Example 66

4.3.4 Results for Comparison 2 70

4.4 Final Conclusions 74

4.5 Future Work 75

5 Appendix **76**

5.1 Tables 76

5.2 Comparison 1 Results 80

5.3 Comparison 2 Results 84

5.4 Glossary 90

5.5 Code 93

6 Bibliography **100**

Chapter 1

Introduction

1.1 The Thesis

There are various testing procedures used in clinical trials with censored data. Some of these procedures are group sequential, analyzing the data at a pre-selected number of equal time intervals and others are pure sequential, where analysis occurs at each event time. The main purpose of sequential analysis is the desire to terminate a study as soon as a decision is possible. For obvious financial and ethical reasons, especially in clinical trials, it is of primary interest to stop a trial as soon as there is an indication that one treatment is superior.

This thesis will analyze and compare different methods in sequential models that test for differences among two treatments. More specifically, the focus will be the comparison of group sequential analysis versus pure sequential analysis. The main parameters of interest for comparison are the power of the test and the required sample size to achieve this power. The power of a test is a measurement of the tests ability to detect differences in populations when one exists. As well, we wish to compare the average stopping times of

the procedures. This is the time (number of observations) needed to make a decision. In a fixed sample setting the stopping time will always be the end of the study, after n observations. Sequential analysis techniques, however, gives us the opportunity of making a decision at an earlier time (the stopping time). The goal is to explore the effectiveness of a pure sequential method versus group sequential models. This is an applied thesis and the results are based on Monte Carlo simulation in C++.

1.2 The Title

1.2.1 Non Parametric...

Non parametric statistical methods are extremely useful in analysis when specific model assumption are in question. In parametric statistical procedure, assumptions about the data are made based on either historical results or some sort of intuition from the analyst. These assumptions are critical when it comes to the decision making process. There are times that we will be confident that any departure from these assumptions will have minimal effect on the outcome so that the resulting conclusions remain valid. In these cases our test is said to be robust to departure from these assumptions. However, when this robustness is lacking or assumptions about the underlying model are in question, we must seek non-parametric methods to answer our problem. Non-parametric methods assume nothing about the distribution of the data in question. Non-parametric methods generally look solely at the ranks of the observations, rather than the observations themselves. In many cases non-parametric methods perform almost as well as parametric methods even when the assumptions are satisfied, and often better when they are not.

1.2.2 ...Sequential Analysis...

In statistical analysis, it is the goal of the statistician to collect sample data and make inferences on the entire population. This is generally done by collecting a sample and analyzing it using some statistical framework. In a fixed sample setting, a fixed number of observations are collected, and a test is performed to reach some conclusion about the population. Sometimes it may be difficult and/or timely to gather all the required data at once. As well, it may be unnecessary. This is of special concern in clinical trials. If a study is testing the effectiveness of drug A versus drug B, we should only run the study just as long as is needed to detect which drug is better, then the superior drug can be administered to all patients. It is not ethical to wait until the end of the study to make a decision, especially when the conclusion could have been made at an earlier time, perhaps saving lives. As well, early termination has its financial benefits as clinical trials can be very expensive. Thus, the need for continuous monitoring, or sequential analysis.

Sequential analysis is the analysis of data periodically throughout the study. Beyond this, sequential analysis can be broken into two categories. The first is pure sequential procedures where an interim analysis is performed on the accumulated data after every single observation becomes available. The second is group sequential methods where interim analysis are performed a relatively small number of time throughout the study, say after every $2N_g$ observations, on the accumulated data. In any case, these methods are of primary interest in clinical trials where it is of special interest to make conclusions as soon as possible for obvious ethical and financial costs.

1.2.3 ...in Clinical Trials with Censored Data.

When dealing with clinical trials, we are often studying the effects of treatments or drugs on live subjects. The goal is to compare the effectiveness of the treatments. This will generally be measured by the survival time of the subjects. Survival time can mean many things, but we will refer to it as the length of time the patient was in the study. When dealing with live subjects, other factors other than the treatment may affect the results. For example, if a patient moved and is unable to continue with the study, we must find a way to deal with the observation. Cases such as this will almost surely happen when dealing with human subjects. Thus, (ignoring the rare occurrence when patients cannot be lost in the middle of the study for reasons beyond the scope of the study), we must find a way to deal with these observations rather then simply ignoring them. These observations are incomplete and are said to be censored. We do not know the exact survival time for censored observations. If a patient exits the study unexpectedly for reasons beyond the scope of the study, for example, the patient moves away and participation becomes impossible, this patient is said to be lost to follow-up. We will call this type I censoring. Another form of censoring occurs if the patient survives to the end of the study and we cannot measure their exact survival time. This observation has type II censoring. In either case we do not know exactly what their survival time is, but only know the patient survived at least the time they spent in the study.

1.3 The Hypothesis Test

In a courtroom trial, someone is accused of committing a crime. When the trial begins it is assumed that the accused is innocent. It is the purpose of the trial to determine if there is enough evidence to convict the accused. To accomplish this, the prosecution must find enough evidence for a conviction, or against the assumed innocence. If there is significant evidence against innocence, the jury would decide in favor of an alternative; a verdict of guilty. In any case, it is the job of the prosecution to find enough evidence against the accused. The process of a courtroom trial is very similar to hypothesis testing in a statistical framework.

One of the fundamental resources in statistics is hypothesis testing. The goal of a hypothesis test is to determine if there is any significant evidence against some assumption; in statistics this assumption is called the null hypothesis. Generally, the null hypothesis is a statement of no difference or no effect. For example, when a researcher develops a drug to treat some disease he/she must convince people of its effectiveness. To do this the new drug must be tested versus another treatment so its relative effectiveness can be measured. To conduct this comparison a hypothesis test will be performed. The hypothesis test will assume there is no difference between the two drugs, this is the null hypothesis (the assumption of no difference). The goal of the researcher is to prove his/her new drug is better. Based of the data collected the test will determine if there is enough evidence to reject the null hypothesis of no treatment difference. If the null hypothesis is rejected the alternative conclusion must be made that there is some difference between the two drugs. In this case, since the researcher wishes to prove his/her drug is better, the

alternative hypothesis will be that the new drug is better. This type of application can be seen in many aspects of normal everyday life. Going back to the courtroom, the presumption of innocence is the null hypothesis. From this, it is the job of the prosecution to negate this assumption and find convincing or significant evidence to warrant a conviction. This is analogous to rejecting the null hypothesis in a statistical point of view. Here the alternative is to find the accused guilty.

Other similarities exist between these examples. One of the most important points to remember is that we are strictly trying to find evidence against the null hypothesis in favor of the alternative. In general we are not trying to find evidence in favor of the null hypothesis. In fact, we never accept the null hypothesis, we can only say there is no evidence against it. This may seem suspicious in the courtroom setting. In a courtroom trial, a jury can find the accused not guilty, but they never find an accused innocent. It is extremely important to distinguish between the two. It is not the job of the defence to prove innocence (this is already assumed), it is only their job to show the evidence against innocence is insufficient to warrant a conviction.

Just because someone is found not guilty does not mean that the person is innocent. Likewise, just because we do not reject the null hypothesis does not mean it is true. In either case there is almost surely some evidence against the null hypothesis. In a drug trial, the results of the two drugs will never be exactly the same. The idea of the statistical test is to determine if this difference is attributed to a true difference in drugs or is it due to some random occurrence. Similarly, as we see in courtroom trial, there will always

evidence against the accused (this is essential to warrant a trial in the first place); it is the goal of the trial to determine whether this evidence exists because the accused is in fact guilty or if there some other explanation for its existence. For example, circumstantial evidence almost surely exists, but does not necessarily warrant a conviction.

In either case, we can never be 100% sure of our results. There is always the chance of error. This can happen in two ways, either the null hypothesis is rejected when it's is in fact true, or it is not rejected when it is false. Similarly, the accused could be found guilty when he/she is innocent, or found not guilty when in fact they are guilty. The former is the primary concern. In a statistical point of view it is very undesirable to detect some difference in treatments when in fact no difference exists. Similarly, we never want to convict an innocent person. It is evident we want the chance of this happening to be as small as possible. This error is called type I error and can be controlled in a statistical setting by defining what we mean by significant evidence. The second type of error, type II error, is when we do not reject the null hypothesis when it is false. Again we would like this to be as small as possible, but it is not as crucial as type I error. We can see this in the courtroom setting where a guilty person goes free. One might say, letting a guilty person go free is not as critical as convicting an innocent person.

1.4 A Statistical View of the Clinical Trial

In clinical trials the most important observation is the length of time a patient lasts in the study from time of entry. Initially upon entry a patient is assigned to a treatment group, and immediately the patient begins the treatment. The length of time from the start of the treatment until the patient leaves the study will be called the *survival time*. The time a patient leaves the study will be called the *event time*. We have what is called *censoring* when a patient has not died at the completion of the study, or is lost to follow-up. Their survival time will only be known to be at least the length of time they were in the study. In other cases, patients may be forced to leave the trial for reasons beyond the scope of the study. Again, their survival time will be known to be at least the length of time they were in the study before they were lost to follow-up. This happens when a patient moves and can no longer participate in the trial (perhaps due to geographical constraints). Thus, when a patient exits the study, they have either died or they were censored.

In general, the study is designed to compare the survival times between treatment groups. In this thesis we will concentrate solely on the two treatment case. We will be comparing the survival distributions of two treatments. Suppose we wish to test the null hypothesis that the unknown survival distributions of two competing treatments are equal, that is,

$$\mathrm{H}_0 : S_A(t) = S_B(t), \text{ all } t \tag{1.1}$$

vs. the alternative that they are somehow different

$$\mathrm{H}_a : S_A(t) \neq S_B(t), \text{ for some } t \in \mathbb{R}. \tag{1.2}$$

$S_y(t)$ is the survival distribution defined as the probability an individual will survive past a certain time, or

$$S_y(t) = \mathrm{P}\{T > t\} = \int_t^{\infty} f(s)ds, t \geq 0, \tag{1.3}$$

When comparing survival data it is common to compare the hazard ratio, λ_h, of the two competing treatments. The hazard ratio is the ratio of the hazard rates of the two survival distributions where the hazard rate is described as a subject's instantaneous chance of failure at time t, given the subject is alive at time t. Hence, the hazard rate is

$$h(t) = \lim_{\Delta \to 0} \frac{\mathrm{P}\left\{\text{failure} \in (t, t+\Delta) \text{ given alive at time } t\right\}}{\Delta} \tag{1.4}$$

or equivalently,

$$h(t) = \frac{f(t)}{S(t)}, t \geq 0. \tag{1.5}$$

A special case of the hypothesis test in (1.1) and (1.2) is a test on the ratio of the hazard rates from two treatments. We will see this more when we discuss the log-rank test. A special case of (1.1) and (1.2) is thus

$$\mathrm{H}_0 : \lambda_h = \frac{h_B(t)}{h_A(t)} = 1 \tag{1.6}$$

vs.

$$\mathrm{H}_a : \lambda_h = \frac{h_B(t)}{h_A(t)} \neq 1 \tag{1.7}$$

It will be assumed patients enter serially throughout the study. This is called *staggered* entry. Upon entry, the subjects are randomly assigned to a

treatment group through some random allocation mechanism. We will only focus on the two sample case. Hence, if the probability a subject is assigned to group A is φ, where $0 < \varphi < 1$, then the probability a subject is assigned to another group B is 1 - φ. We will generally take $\varphi = 0.5$. We will denote k as the total number of patients in the trial at time of analysis, where $k = k_A + k_B$. k_A is the number of patients receiving treatment A and k_B is the number of patients receiving treatment B. k_A and k_B are random variables, and by the assignment rule,

$$\mathrm{E}[k_A] = \varphi k \tag{1.8}$$

and

$$\mathrm{E}[k_B] = (1 - \varphi)k \tag{1.9}$$

As well, for future notation, let

$$\pi_j = \begin{cases} 1 & \text{if the } j\text{-th subject belongs to group } A; \\ 0 & \text{if the } j\text{-th subject belongs to group } B. \end{cases} \tag{1.10}$$

Upon entry the subject's arrival time is recorded as the time since the beginning of the study. We will denote the arrival time of subject j assigned to treatment y as $U_j^{(y)}$, where $y = A, B, j = 1, 2, \ldots$. Upon assignment to a treatment group the subject $\{j, y\}$ will immediately begin the treatment. This subject will continue the treatment until they have died, they are lost to-follow up or the study ends. If this subject dies their **exact** survival time will be denoted as $T_j^{(y)}$. If a patient is lost to follow-up or survives the study (doesn't die) their survival time is censored and will be denoted as $C_j^{(y)}$. Let $X_j^{(y)} = \min\left\{T_j^{(y)}, C_j^{(y)}\right\}$, denote the patient's survival time, or simply the observed time spent in the study.

Let $E_j^{(y)}$ denote the event time, or the time the patient left the study, so that $E_j^{(y)} = U_j^{(y)} + X_j^{(y)}$. We will denote $\Delta_j^{(y)}$ as the censoring variable so that,

$$\Delta_j^{(y)} = \begin{cases} 1 & \text{if } X_j^{(y)} = C_j^{(y)}; \\ 0 & \text{if } X_j^{(y)} = T_i^{(y)}. \end{cases} \tag{1.11}$$

We will assume the distribution of Δ is identical for both treatment groups, that is, $P\{\Delta = 1\}$ is constant for all subjects in the study. As well, it is feasible to assume continuous distribution so that at time k we will observe distinct event times $E_{(1)} < \ldots < E_{(k)}$.

Almost surely, even under H_0, the observed survival times will be different between the two treatment groups. Basic sample statistics can be calculated to see the estimated difference. However, the question is whether this difference is due to a real difference in the population survival times or is it due to mere chanc e variation within the entire population? Statistical methods address this question. The purpose of this thesis is to compare the results of a group sequential procedure to those of a pure sequential procedure. In either case, interim analyses will be conducted throughout the study.

In the group sequential methods discussed, a fixed number of analyses K will be conducted, each after a required number of patients have died. In each analysis, a statistic will be calculated and tested for its significance. If at any time the results are found to be statistically significant, the study will end and the null hypothesis will be rejected. The null hypothesis is not rejected if no significant evidence is found at the end of the K-th analysis.

In the case of the pure sequential method, an analysis is performed after every event or after each patient leaves the study. The null hypothesis will be rejected the first time we find significant evidence of a difference between the treatments. Otherwise, if no significant evidence is found at the end of the study, the null hypothesis will not be rejected. In any case, the format of the test is virtually identical.

In Chapter 2, we will review the non-parametric tests we will be exploring as well as the basic rank statistics used. We will review the group sequential U-statistics and their application with the Wilcoxon and Gehan scores. Then we will review Mantel's (1966) fixed sample Log-Rank test, followed by it's extension to group sequential analysis using joint distribution theory proposed in Jennison and Turnbull (1999). Finally, we will look at a pure sequential design by Gombay (2001) using a U-statistic with the Gehan anti-symmetric kernel.

Chapter 2

Literature Review

This chapter will review some basic theory which is used in the non-parametric settings. Of primary interest is the Wilcoxon and Gehan U-statistics. To address a group sequential test in this thesis the sequential Log-Rank test will be used. I will review the log-rank test in a fixed sample setting as well as introduce its extension to the group sequential setting.

2.1 U-Statistics

Consider a sequence $X_1, \ldots, X_n$ of n independent identically distributed random variables with unknown distribution function $F(x)$. Suppose on the basis of the sample we wish to estimate the population characteristic $\theta(F)$. Let $m \leq n$ be a number and $\psi_m(X_1, \ldots, X_m)$ be a function of m variables. Define

$$U_m = \binom{n}{m}^{-1} \sum_{(n,m)} \psi_m(x_{i_1}, \ldots, x_{i_m}), \tag{2.1}$$

where the sum is over all m-degree combinations of $X_1, \ldots, X_m$. U_m is an unbiased estimator of θ(F) and is called a U-statistic (Hoeffding, 1948).

2.2 Wilcoxon Type Statistics

This section will review the Wilcoxon rank statistic and its expansion to the Mann-Whitney U-statistic allowing for tied observation, finishing off with the Gehan statistic which allows for tied and censored observations. The description is in terms of a fixed sample test with n observations.

For the purpose of this paper we are dealing with survival data, and testing on survival distributions. Recall, for the test

$$\mathrm{H}_0 : S_A(t) = S_B(t), \text{ all } t \tag{2.2}$$

versus the alternative that they are different, that is

$$\mathrm{H}_a : S_A(t) \neq S_B(t), \text{ for some } t. \tag{2.3}$$

Assume we have a total of n observation from two populations, A and B. Earlier we denoted survival times by $X_j^{(y)} = \min\left\{T_j^{(y)}, C_j^{(y)}\right\}$, $y = A, B, j =$

$1, 2, \ldots, n$. Recall we are assuming some continuous distribution, and thus we will have no ties.

First, consider data without censoring, so $X_j^{(y)} = T_j^{(y)}$. Combining the sequence of survival times $T_j^{(y)}$ from the two populations, we have $X_{(1)} < \ldots < X_{(n)}$ ordered observations, where $n = n_A + n_B$. Let R_j denote the rank of the j-th ordered observation. Let

$$W_n = \sum_{j=1}^{n} R_j \pi_j. \tag{2.4}$$

Recall,

$$\pi_j = \begin{cases} 1 & \text{if the j-th subject belongs to group A;} \\ 0 & \text{if the j-th subject belongs to group B.} \end{cases} \tag{2.5}$$

Then, W_n is the sum of the ranks from observation in group A.

Using the large-sample approximation, W_n is asymptotically normal. Under the null hypothesis (2.2),

$$\mathrm{E}[W_n] = \frac{n_A(n+1)}{2} \tag{2.6}$$

and

$$\mathrm{Var}[W_n] = \frac{n_A n_B(n+1)}{12} \tag{2.7}$$

Therefore, standardizing W_n, we have,

$$Z_W = \frac{W_n - \mathrm{E}[W_n]}{\sqrt{\mathrm{Var}[W_n]}} \sim \mathrm{N}(0, 1) \tag{2.8}$$

Z_W is the Wilcoxon rank sum standardized statistic, and the subscript W defines it as such.

Since, Z_W follows approximately a standard normal distribution if n is large, for the two-sided alternative hypothesis (2.3), we reject the null hypothesis (2.2) at significance level alpha if $|Z_W| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the standard normal cumulative distribution function.

2.2.1 Mann-Whitney U statistics

Consider the test statistic

$$U_n = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} h\left(X_i^{(A)}, X_j^{(B)}\right) \tag{2.9}$$

where

$$h\left(X_i^{(A)}, X_j^{(B)}\right) = \begin{cases} 1 & \text{if } X_i^{(A)} > X_j^{(B)}; \\ -1 & \text{if } X_i^{(A)} < X_j^{(B)}. \end{cases} \tag{2.10}$$

It can be shown that

$$W_n = \frac{U_n}{2} + \frac{n_A(n+1)}{2} = \frac{U_n}{2} + \mathrm{E}[W_n] \tag{2.11}$$

Tests based on W_n or U_n are equivalent in the sense that their standardized values are identical. The statistic U_n, with kernel function (2.10) is called the Mann-Whitney U-statistic.

This statistic can be generalized to accommodate the case where tied observation may arise. Let

$$h\left(X_i^{(A)}, X_j^{(B)}\right) = \begin{cases} 1 & \text{if } X_i^{(A)} > X_j^{(B)}; \\ 0 & \text{if } X_i^{(A)} = X_j^{(B)}; \\ -1 & \text{if } X_i^{(A)} < X_j^{(B)}. \end{cases} \tag{2.12}$$

and

$$U_n = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} h(X_i^{(A)}, X_j^{(B)}). \tag{2.13}$$

In the case of tied observation, the relationship between U_n and W_n still remains and it can be shown that again

$$W_n = \frac{U_n}{2} + \frac{n_A(n+1)}{2} = \frac{U_n}{2} + \mathrm{E}[W_n]. \tag{2.14}$$

Using large sample approximation we can standardize the test statistic W_n or U_n to obtain a random variable which follows an approximate standard normal distribution. Using U_n with kernel function (2.12) we have,

$$Z_{MW} = \frac{U_n - \mathrm{E}[U_n]}{\sqrt{\mathrm{Var}[U_n]}} \sim \mathrm{N}(0,1). \tag{2.15}$$

To calculate $\mathrm{E}[U_n]$ and $\mathrm{Var}[U_n]$, we refer to equation (2.14). Note that, in the case of tied observations, W_n is still the sum of the ranks for observations from group A. The only difference is that now we use average ranks in the case of ties. In the complete data set (no censoring) the distribution of W_n under the null hypothesis remains the same and (2.6) and (2.7) hold. From this it is easily shown that in (2.14),

$$U_n = 2 \times (W_n - n_A(n+1)) \tag{2.16}$$

so that

$$\mathrm{Var}[U_n] = 4 \times \mathrm{Var}[W_n] \tag{2.17}$$

and from (2.6) and (2.7) we have,

$$\mathrm{E}[U_n] = 2 \times \mathrm{E}[W_n] - n_A(n+1) = 0, \tag{2.18}$$

and

$$\text{Var}[U_n] = 4 \times \text{Var}[W_n] = \frac{n_A n_B (n+1)}{3}. \tag{2.19}$$

Since Z_{MW} follows a standard normal distribution, for the two-sided alternative hypothesis (2.3), we reject the null hypothesis (2.2) at significance level alpha if $|Z_{MW}| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the standard normal cumulative distribution function.

2.2.2 Gehan Test for censored data

Now, consider the case where we have censored data. This result is the most general case and can be reduced to the prior results in (2.10) and (2.12) where censoring or tied observations do not exist. When dealing with censored data there are times when we cannot be sure if an observation is larger than another. Here we assume that the probability of a censored observation is the same for all patients in either group. Gehan generalizes the result in (2.12) to accommodate for censored observation, ties still permitted. Referring back to our general notation, our observations are defined as follows, $X_j^{(y)} = \min\left\{T_j^{(y)}, C_j^{(y)}\right\}$. And for $j = 1, \ldots, n$ and $y = A, B$,

$$\Delta_j^{(y)} = \begin{cases} 1 & \text{if } X_j^{(y)} = C_j^{(y)}; \\ 0 & \text{if } X_j^{(y)} = T_j^{(y)}. \end{cases} \tag{2.20}$$

Now, consider the sequence $\left\{X_j^{(y)}, \Delta_j^{(y)}\right\}$, where $n = n_A + n_B$, $y = A, B$, and $j = 1, \ldots, n$.

Gehan (1965) proposed the following generalization of the Wilcoxon score for censored data,

$$h\left(X_i^{(A)}, X_j^{(B)}\right) = \begin{cases} 1 & \text{if } X_i^{(A)} > X_j^{(B)}, \Delta_j^{(B)} = 0; \\ 1 & \text{if } X_i^{(A)} = X_j^{(B)}, \Delta_i^{(A)} = 1, \Delta_j^{(B)} = 0; \\ 0 & \text{otherwise;} \\ -1 & \text{if } X_i^{(A)} < X_j^{(B)}, \Delta_i^{(A)} = 0; \\ -1 & \text{if } X_i^{(A)} = X_j^{(B)}, \Delta_i^{(A)} = 0, \Delta_j^{(B)} = 1. \end{cases} \tag{2.21}$$

and again we use

$$U_n = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} h\left(X_i^{(A)}, X_j^{(B)}\right). \tag{2.22}$$

In the case of censored data, we must consider a more specific null hypothesis,

$$\mathrm{H}_0^* : S_{T_A}(t) = S_{T_B}(t) \text{ and } S_{C_A}(t) = S_{C_B}(t), \text{for all t,} \tag{2.23}$$

where $S_{T_y}(t)$ and $S_{C_y}(t)$ are the distributions of the uncensored and censored observations in group y, respectively. If no censoring occurs, the mean and variance of the Gehan U-statistic reduces exactly to the Mann-Whitney U-statistic. However, when censoring exists, the variance of the U-statistic isn't easily calculated using Gehan's score. As well, for large sample sizes, Gehan's statistic itself can be very tedious to calculate. Mantel (1967) shows a much simpler calculation which scores each observation with respect to its rank in the combined sample. As well he shows the calculation of the variance is much simpler.

Consider the combined sample, $\left\{X_1^A, \Delta_1^A\right\}, \ldots, \left\{X_{n_A}^A, \Delta_{n_A}^A\right\}, \left\{X_1^B, \Delta_1^B\right\}, \ldots, \left\{X_{n_B}^B, \Delta_{n_B}^B\right\}$. Redefine the sample to be $\{Z_1, \Delta_1\}, \ldots, \{Z_n, \Delta_n\}$. Let's reconsider expression (2.21). For $m = 1, \ldots, n$ and $l = 1, \ldots, n$,

$$h\left(Z_m, Z_l\right) = \begin{cases} 1 & \text{if } Z_m > Z_l, \Delta_l = 0; \\ 1 & \text{if } Z_m = Z_l, \Delta_m = 1, \Delta_l = 0; \\ 0 & \text{otherwise;} \\ -1 & \text{if } Z_m < Z_l, \Delta_m = 0; \\ -1 & \text{if } Z_m = Z_l, \Delta_m = 0, \Delta_l = 1. \end{cases} \tag{2.24}$$

Consider an $n \times n$ symmetric matrix where the entry in row k and column l is defined as $h\left(Z_m, Z_l\right)$. Summing over the rows, we have,

$$U_m^* = \sum_{l=1; l \neq m}^{n} h\left(Z_m, Z_l\right). \tag{2.25}$$

The sum of the values in row m is precisely the number of remaining n-1 observations that observation Z_m is surely greater than minus the number of remaining observations that observation Z_m is surely less than. Summing these U_m^* only for group 1 we achieve the same U-statistic as Gehan.

$$U_n = \sum_{m=1}^{n} U_m^* \pi_m. \tag{2.26}$$

where π_m is defined in (2.5).

To calculate the distribution of U_n, consider the permutational variance proposed by Mantel (1967). This can be derived by considering all possible ways of selecting n_A observations from the n observations. The test statistic U_n, in this case, is asymptotically normal with mean zero and,

$$\text{Var}(U_n) = \frac{n_A n_B}{n(n-1)} \sum_{i=1}^{n} \left(U_i^*\right)^2 \tag{2.27}$$

Using large sample approximations we have,

$$Z_G = \frac{U_n}{\sqrt{\mathrm{Var}(U_n)}} \sim \mathrm{N}(0,1). \tag{2.28}$$

Since Z_G follows a standard normal distribution, for the two-sided alternative hypothesis (2.3), we reject the null hypothesis (2.2) at significance level alpha if $|Z_G| > Z_{1-\alpha/2}$, where $Z_{1-\alpha/2}$ is the standard normal cumulative distribution function.

2.3 The Log-Rank test

Suppose we want to test the equality of the survival distributions

$$\mathrm{H}_0 : S_A(t) = S_B(t), \text{ for all } t \tag{2.29}$$

vs. the alternative that they are somehow different

$$\mathrm{H}_a : S_A(t) \neq S_B(t), \text{ for some } t \in \mathbb{R} \tag{2.30}$$

or equivalently in a hazard ratio test

$$\mathrm{H}_0 : \theta = \log(\lambda_h) = \log\left(\frac{h_B(t)}{h_A(t)}\right) = 0, \text{ for all } t \tag{2.31}$$

vs.

$$\mathrm{H}_a : \theta = \log(\lambda_h) = \log\left(\frac{h_B(t)}{h_A(t)}\right) \neq 0 \text{ for some } t \in \mathbb{R} \tag{2.32}$$

Speaking in terms of a fixed sample test ($K = 1$) there is only 1 analysis which occurs at the end of the study. The following description is in terms of multiple testing which is done sequentially throughout the study. For any given analysis, the log-rank statistic is calculated just as it is in the fixed sample setting using the accumulated data up to that analysis. This is repeated significance testing or group sequential testing. In either case we want the overall type I error to remain the same, thus the extension from a fixed sample case to a sequential environment is as simple as choosing the appropriate significance level at each test to ensure the overall significance level (the probability of rejecting H_0 when it is true at any analysis) remains the same. We will discuss

the theory behind this later. For now, in terms of the fixed sample setting, simply regard $K = 1$ in the following description.

Suppose at analysis k' we observe a total number of deaths (uncensored observations), $N_{k'}$, $k' = 1, \ldots, K$. Assuming continuous survival distributions we will have $N_{k'}$ distinct survival times for these subjects, $X_{(1),k'} < \ldots < X_{(N_{k'}),k'}$. Then, at any time j, $j = 1, \ldots, N_{k'}$ at analysis k', we will have $r_{j,k'}$ surviving patients, where $r_{j,k'} = r_{jA,k'} + r_{jB,k'}$.

A non-parametric statistic, the log-rank statistic is calculated as

$$LR_{k'} = \sum_{j=1}^{N_{k'}} \left\{ \pi_{jA,k'} - \frac{r_{jA,k'}}{r_{j,k'}} \right\}, \tag{2.33}$$

where

$$\pi_{jA,k'} = \begin{cases} 1 & \text{if the failure at time j was in treatment A;} \\ 0 & \text{if the failure at time j was in treatment B.} \end{cases} \tag{2.34}$$

The estimated variance, at analysis k', for $LR_{k'}$ is

$$\widehat{\text{Var}}(LR_{k'}) = \sum_{i=1}^{N_{k'}} \frac{r_{iA,k'} \times r_{iB,k'}}{r_{i,k'}^2} \tag{2.35}$$

Under the null hypothesis, at any time, we'd expect that the number at risk in each treatment to be the same, so that $r_{jA,k'} = r_{jB,k'} = \frac{r_{j,k'}}{2}$. So under the null hypothesis, at analysis k', the expected value of the log-rank statistic is,

$$\text{E}\left[\sum_{j=1}^{N_{k'}} \pi_{jA,k'}\right] - \text{E}\left[\sum_{j=1}^{N_{k'}} \frac{r_{jA,k'}}{r_{j,k'}}\right] = \frac{N_{k'}}{2} - \frac{N_{k'}}{2} = 0. \tag{2.36}$$

Broken down, $\pi_{jA,k'}$ is the observed number of failures in group A at time j, and $r_{jA,k'}/r_{j,k'}$ is the expected number of failures in group A to occur at time j. As well, the variance under the null hypothesis is,

$$\sum_{i=1}^{N_{k'}} \frac{r_{iA,k'} \times r_{iB,k'}}{r_{i,k'}^2} = \sum_{i=1}^{N_{k'}} \frac{r_{i,k'} \times r_{i,k'}}{4r_{i,k'}^2} = \sum_{i=1}^{N_{k'}} \frac{1}{4} = \frac{N_{k'}}{4}. \tag{2.37}$$

Standardizing the log-rank statistic we have, under H_0,

$$Z_{k'} = \frac{LR_{k'}}{\sqrt{\widehat{\text{Var}}(LR_{k'})}} \sim \text{N}(0,1), k' = 1, \ldots, K. \tag{2.38}$$

In the fixed sample setting with $K = 1$ (and $k' = 1$ at the only analysis), a decision is made at the end of the study by comparing the test statistic Z_1 to a critical value c defined under the probability model of its null distribution $Z \sim N(0,1)$ so that $P\{|Z_1| > c\} = \alpha$. We will only reject the null hypothesis in (2.29) if $|Z_1| > c$ or equivalently $2P\{Z > |Z_1|\} < \alpha$.

We will extend this to a group sequential environment by defining nominal type I errors for each analysis, k', where $k' = 1, \ldots, K$, so that the overall type I error is fixed at α, which is the probability of rejecting the null hypothesis at any time during the study when it is in fact true. To do this we must choose a sequence of K interim critical values, $\{c_1, \ldots, c_K\}$. These values are calculated according to the chosen nominal significance levels, $\alpha_1, \ldots, \alpha_K$, so that the probability under the joint distribution of the sequence of K test statistics is such that

$$\text{P}\{|Z_1| > c_1, \text{or}, \ldots, \text{or}, |Z_K| > c_K\} = \alpha \tag{2.39}$$

where

$$2\mathrm{P}\{|Z_{k'}| > c_{k'}\} = \alpha_{k'} \tag{2.40}$$

for $k' = 1, \ldots, K$.

Choosing the sequence of critical values is not unique. Pocock (1977) offers a sequence of critical values so that the nominal significance levels are constant for each analysis, where O'Brien and Fleming (1979) decide to be more conservative at first, with smaller nominal significance levels, leaving more opportunity to detect a difference in later stages. In either case, the overall type I error is the same.

Chapter 3

Sequential Designs

Here a general approach to group sequential testing will be reviewed with the special case where patient accrual must be equal for each analysis. This group sequential setting with equal groups will incorporate the error distribution theory from Pocock (1977), and O'Brien and Fleming (1979). Examples of the general group sequential approach will follow, most interesting its application to the log-rank test. Last, we will review a purely sequential method using the U-statistic proposed by Gombay (2001). All methods discussed in this thesis are two-sample problems with censored data.

3.1 A General Approach to Group Sequential Testing

Before we begin looking at the decision making approaches proposed by Pocock (1977), and O'Brien and Fleming(1979), a general framework will be presented to enable the use of such theories to any sequence of statistics. In this paper we use these results to generalize the fixed sample log-rank test to a group sequential approach. It should be noted that many other statistics are equally applicable under the properties of the following group sequential test statistics. A few examples are included.

Suppose in a group sequential study, a maximum of K interim analyses will be performed. For each analysis a test is performed to determine whether early stopping can occur. A test statistic is calculated at each analysis based on the accumulated data. From this we obtain the sequence, $\{Z_1, \ldots, Z_K\}$, of standardized test statistics. For each analysis we calculate these statistics based on the accumulated data. As we proceed through the study our sample size gets larger, and thus our information gets stronger. For some population parameter θ, these standardized statistics are said to follow the canonical joint distribution, Jennison and Turnbull (1999), with accumulated information levels $\{I_1, \ldots, I_K\}$ at each analysis if:

1. $(Z_1, \ldots, Z_K)$ is multivariate normal,
2. $\mathrm{E}[Z_{k'}] = \theta\sqrt{I_{k'}}, k' = 1, \ldots, K$ and,
3. $\mathrm{Cov}[Z_{k'_1}, Z_{k'_2}] = \sqrt{k'_1/k'_2}, 1 \leq k'_1 \leq k'_2 \leq K.$

This is the conditional distribution of the sequence of standardized test statistics, $\{Z_1, \ldots, Z_K\}$ given $\{I_1, \ldots, I_K\}$. Note that $\mathrm{Var}[Z_{k'}] = \mathrm{Cov}[Z_{k'}, Z_{k'}] = 1$

so that

$$Z_{k'} \sim \mathrm{N}\left(\theta\sqrt{I_{k'}}, 1\right). \tag{3.1}$$

Jennison and Turnbull (1999) also offer the joint canonical distribution in terms of other statistics. Consider the data $Y_j \sim \mathrm{N}(\theta, 1)$, $i = 1, 2, \ldots,$ with $N_{k'}$ observation at analysis k', $k' = 1, \ldots, K$. At analysis k', we calculate the score statistic $S_{k'} = Y_1 + \ldots + Y_{N_{k'}}$. The sequence $\{S_1, \ldots, S_{k'}\}$ is then multivariate normal with

$$S_{k'} \sim \mathrm{N}(\theta I_{k'}, I_{k'}) \text{ and } \mathrm{Cov}[S_i, S_j] = I_i \tag{3.2}$$

for $k' = 1, \ldots, K$ and $1 \leq i \leq j \leq K$. This sequence of statistics also follows the joint canonical distribution.

3.1.1 Tests With Equal Group Sizes

In general, suppose we wish to test the null hypothesis $\mathrm{H}_0 : \theta = 0$ versus $\mathrm{H}_a : \theta = \pm\delta$, for some unknown population parameter θ. Suppose also that it is required the type I error or significance level be set at α and at $\theta = \pm\delta$ the power be $1 - \beta$. We consider a maximum of K analyses and at each analysis a standardized test statistics is calculated and the sequence of test statistics $\{Z_1, \ldots, Z_K\}$ are observed. Suppose these test statistics follow the joint canonical distribution, defined in section 3.1, with information levels $\{I_1, \ldots, I_K\}$ for estimating θ. The methods we will study later, (Pocock (1977), and O'Brien and Fleming (1979)), require equal group sizes for each analysis, which produces equally spaced information levels. The goal is to determine the information levels required to achieve the pre-determined type I error and power at a

given δ. Thus, $I_{k'}$ will depend on K, k', α, β, δ and the type of significance level approach.

In the fixed sample setting our test statistic Z_f follows a $\mathrm{N}\left(\theta\sqrt{I}, 1\right)$. The letter **f** denotes the fixed sample test. Thus $Z_f \sim \mathrm{N}(0,1)$ under the null hypothesis. We set our two-sided significance level or type I error by rejecting the null hypothesis only if $|Z_f| \geq Z_{1-\alpha/2}$ where $Z_{1-\alpha/2}$ is the standard normal cumulative distribution function so that for $Z \sim \mathrm{N}(0,1)$. In any fixed sample Z test we can calculate the required sample needed to satisfy the power requirement at significance level α. For a simple fixed two sample test with n observations from each sample, we need,

$$\alpha/2 = \mathrm{P}\left\{\hat{\theta} \geq c\right\} = \mathrm{P}\left\{\frac{\hat{\theta}}{\sqrt{2\sigma^2/n}} \geq \frac{c}{\sqrt{2\sigma^2/n}}\right\} = \mathrm{P}\left\{Z_f \geq Z_{1-\alpha/2}\right\}. \quad (3.3)$$

$$\beta = \mathrm{P}\left\{\hat{\theta} \leq c\right\} = \mathrm{P}\left\{\frac{\hat{\theta} - \delta}{\sqrt{2\sigma^2/n}} \leq \frac{c - \delta}{\sqrt{2\sigma^2/n}}\right\} = \mathrm{P}\left\{Z_f \leq -Z_{1-\beta}\right\}. \quad (3.4)$$

when $\theta = \pm\delta$. Therefore we find that,

$$n = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2}. \quad (3.5)$$

Recall, from section 3.1 $\mathrm{E}[Z_f] = \theta\sqrt{I_f}$. We find that,

$$I_f = \frac{\{Z_{1-\alpha/2} + Z_{1-\beta}\}^2}{\delta^2}. \quad (3.6)$$

This is the total required information for the fixed sample test required to

achieve the power $1-\beta$ at $\theta = \pm\delta$.

In group sequential procedures, a larger maximum sample size is required, Jennison and Turnbull (1999). This results in a larger maximum information level, which we will set to $I_g = RI_f$, where I_g is the maximum information needed in a group sequential test and $R > 1$. $R = R(K, \alpha, \beta)$ is a multiplier that also depends on the significance level approach taken. For equal group sizes, at each of the K groups, we will have equally spaced information levels so that the information at analysis k' is,

$$I_{k'} = \frac{k'}{K} I_g = \frac{k'R}{K} I_f, k' = 1, \ldots, K. \tag{3.7}$$

Letting

$$M = \frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\sqrt{K}} \tag{3.8}$$

the information at analysis k' can be re-written as

$$I_{k'} = \frac{k'R}{\delta^2} \times M^2, k' = 1, \ldots, K. \tag{3.9}$$

Therefore, following our general framework of the joint canonical distribution described earlier, under $H_0 : \theta = 0$ we have,

1. $(Z_1, \ldots, Z_K)$ is multivariate normal,

2. $E[Z_{k'}] = 0$ and,

3. $\text{Cov}[Z_{k'_1}, Z_{k'_2}] = \sqrt{\frac{k'_1 RM^2}{\delta^2} / \frac{k'_2 RM^2}{\delta^2}} = \sqrt{k'_1/k'_2}, 1 \leq k'_1 \leq k'_2 \leq K.$

Under $H_a : \theta = \pm\delta$, we have,

1. $(Z_1, \ldots, Z_K)$ is multivariate normal,

2. $\mathrm{E}[Z_{k'}] = \pm M\sqrt{k'R}, k' = 1, \ldots, K$ and,

3. $\mathrm{Cov}[Z_{k'_1}, Z_{k'_2}] = \sqrt{k'_1/k'_2}, 1 \leq k'_1 \leq k'_2 \leq K.$

If after analysis k', $k' = 1, \ldots, K-1$, $Z_{k'}$ is found to be significant, that is

$$|Z_{k'}| \geq c_{k'}, \tag{3.10}$$

then we reject the null hypothesis and stop the study, otherwise we continue on to analysis $k'+1$. If at analysis K, Z_K is found to be significant,

$$|Z_K| \geq c_K, \tag{3.11}$$

then we reject H_0, otherwise we don't reject H_0 and conclude there is no statistical evidence against the null hypothesis. This is the general framework we will be using for group sequential procedures.

3.1.2 Normal Responses

Suppose X_{Ai} and X_{Bi}, $i = 1, \ldots, KN_g$ are the responses of subjects from two treatment groups A and B for K interim analysis with N_g uncensored observations from each treatment group for each analysis. Suppose the observations are such that $X_{Ai} \sim \mathrm{N}(\mu_A, \sigma^2)$ and $X_{Bi} \sim \mathrm{N}(\mu_B, \sigma^2)$, $i = 1, \ldots, KN_g$. Let $\bar{X}_{Ak'}$ and $\bar{X}_{Bk'}$ denote the average response up to analysis k' for each treatment group. Suppose we wish to test the null hypothesis, $\mathrm{H}_0 : S_A(t) = S_B(t)$. Assuming σ^2 is the same for both populations, this is equivalent to testing, $\mathrm{H}_0 : \mu_A - \mu_B = 0$ versus the alternative, $\mathrm{H}_a : \mu_A - \mu_B = \pm\delta$. At analysis k'

$$\hat{\mu}_A - \hat{\mu}_B = \bar{X}_{Ak'} - \bar{X}_{Bk'}. \tag{3.12}$$

The distribution of the estimate is defined as,

$$\bar{X}_{Ak'} - \bar{X}_{Bk'} \sim \mathrm{N}\left(\mu_A - \mu_B, \frac{2\sigma^2}{k'N_g}\right). \tag{3.13}$$

Standardizing, we have

$$Z_{k'} = \frac{\bar{X}_{Ak'} - \bar{X}_{Bk'}}{\sqrt{2\sigma^2/k'N_g}} \sim \mathrm{N}\left((\mu_A - \mu_B)\sqrt{k'N_g/2\sigma^2}, 1\right) \tag{3.14}$$

The sequence $\{Z_1, \ldots, Z_K\}$ is multivariate normal since the marginal distribution of each $Z_{k'}$ is a linear combination of independent normal observations, $\bar{X}_{Ai}$ and $\bar{X}_{Bi}$, for $i = 1, 2 \ldots, KN_g$. Jennison and Turnbull (1999) show that the $\mathrm{Cov}\left[Z_{k'_1}, Z_{k'_2}\right] = \sqrt{k'_1/k'_2}$. Therefore, the sequence of statistics $\{Z_1, \ldots, Z_K\}$ follows the joint canonical distribution in section 3.1 with information levels $I_{k'} = k'N_g/2\sigma^2$, $k' = 1, \ldots, K$, and $\theta = \mu_A - \mu_B$. Under the null hypothesis, $Z_{k'} \sim \mathrm{N}(0, 1)$. If we wish to test H_0 with significance level α and power $1 - \beta$ at $\mu_A - \mu_B = \pm\delta$ our information level I is,

$$I_{k'} = \frac{k'RM^2}{\delta^2}. \tag{3.15}$$

Equating (3.15) with $I_{k'}$ under normal responses ($I_{k'} = k'N_g/2\sigma^2$),

$$N_g = \frac{2RM^2\sigma^2}{\delta^2} \tag{3.16}$$

is the number of observed deaths for each treatment for each analysis. Values for R can be found in the appendix for Pocock, and O'Brien and Flemings models. Their calculations are not found in this thesis but can be found in Jennison and Turnbull (1999). As well values for M have been calculated for different combinations of α and β.

In a fixed sample test, $R = 1$ and $K = 1$. If we wish to test H_0 with significance level α and power $1-\beta$ at $\mu_A - \mu_B = \pm\delta$ our information level I_f is,

$$I_f = \frac{M^2}{\delta^2} = \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{\delta^2} \tag{3.17}$$

From this our required sample size is

$$n_f = 2\sigma^2 I_f. \tag{3.18}$$

3.1.3 Exponential Responses

The section on normal responses was introduced only as a preliminary example. In clinical trials survival times often follow an exponential distribution. Thus it is of interest to set up our hypothesis in terms of the mean difference in exponential observations. Again suppose X_{Ai} and X_{Bi}, $i = 1, \ldots, KN_g$ are the responses of subjects from two treatment groups A and B for K interim analyses with N_g uncensored observations from each treatment group for each analysis. Now suppose the observation are such that $X_{Ai} \sim \text{Exp}(\lambda_A)$ and

$X_{Bi} \sim \text{Exp}(\lambda_B)$, $i = 1, \ldots, KN_g$. Recall, if $Y \sim \text{Exp}(\lambda)$ then the distribution of Y is defined as,

$$f(y) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x}, \tag{3.19}$$

and has the following properties,

$$\text{E}[Y] = \lambda \tag{3.20}$$

and

$$\text{Var}[Y] = \lambda^2. \tag{3.21}$$

Let $\bar{X}_{Ak'}$ and $\bar{X}_{Bk'}$ denote the average response up to analysis k', for $k' = 1, \ldots, K$, for each treatment group. Since the exponential distribution is determined by only one parameter, testing for equal distributions is equivalent to testing the null hypothesis that $\text{H}_0 : \lambda_A - \lambda_B = 0$ versus the alternative $\text{H}_a : \lambda_A - \lambda_B = \pm\delta$. This is not as straight forward as it is in the normal case. Here, an additive effect (δ) is not feasible due to the fact that the variability of the responses depends on the mean. Thus $\lambda_A - \lambda_B$ could be constant, but depending on the magnitude of the expected values, the variances will be completely different. For example, if $\lambda_A = 1$ and $\lambda_B = 2$ it will be a lot easier to detect a difference (more power) than if $\lambda_A = 100$ and $\lambda_B = 101$. In either case though, we have $\delta = 1$. To remedy this, we will look at the difference in their logged values and test, $\text{H}_0 : \log\lambda_A - \log\lambda_B = 0$ versus $\text{H}_a : \log\lambda_A - \log\lambda_B = \delta$ or equivalently $\text{H}_0 : \log\left(\frac{\lambda_A}{\lambda_B}\right) = 0$ versus $\text{H}_a : \log\left(\frac{\lambda_A}{\lambda_B}\right) = \delta$. Our estimate at analysis k' is thus $\log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right)$. Using the delta method and the central limit theorem, it can be shown that,

$$\log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right) \approx \mathrm{N}\left(\log\left(\frac{\lambda_A}{\lambda_B}\right), \frac{2}{k'N_g}\right). \tag{3.22}$$

This was mentioned in Pocock(1977). Our standardized test statistics are,

$$Z_{k'} = \log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right) \times \sqrt{k'N_g/2} \sim \mathrm{N}\left(\log\left(\frac{\lambda_A}{\lambda_B}\right)\sqrt{k'N_g/2}, 1\right) \tag{3.23}$$

Similarly to section 3.1.2, the sequence of standardized statistics is multivariate normal with $\mathrm{Cov}[Z_{k'_1}, Z_{k'_2}] = \sqrt{k'_1/k'_2}$, $1 \leq k'_1 \leq k'_2 \leq K$, these test statistics follow the joint canonical distribution described in section 3.1. Thus,

$$I_{k'} = \frac{k'N_g}{2}, \tag{3.24}$$

which is the same as in the normal means model with $\sigma^2 = 1$. Similar to the normal case, if we wish to test H_0 with significance level α and at $\log(\lambda_A/\lambda_B) = \pm\delta$ the power is $1-\beta$ our information level I at analysis k', for $k' = 1, \ldots, K$, is

$$I_{k'} = \frac{k'RM^2}{\delta^2}. \tag{3.25}$$

Equating (3.24) and (3.25),

$$N_g = \frac{2RM^2}{\delta^2} \tag{3.26}$$

is the number of observed deaths needed for each treatment for each analy-

sis. Again, values for R can be found in the appendix for Pocock (1977), and O'Brien and Fleming (1979).

3.1.4 Weibull Responses

Another popular model for survival times is the Weibull distribution. Again suppose X_{Ai} and X_{Bi}, $i = 1, \ldots, KN_g$ are the responses of subjects from two treatment groups A and B for K interim analysis with N_g uncensored observations from each treatment group for each analysis. Now suppose the observation are such that $X_{Ai} \sim \text{Weibull}(\lambda_A, \gamma)$ and $X_{Bi} \sim \text{Weibull}(\lambda_B, \gamma)$, $i = 1, \ldots, KN_g$. If $Y \sim \text{Weibull}(\lambda, \gamma)$ then Y is described by the two parameter function,

$$f(y) = \left(\frac{1}{\lambda}\right)^{\gamma} \gamma y^{\gamma-1} e^{-\left(\frac{x}{\lambda}\right)^{\gamma}} \tag{3.27}$$

with

$$\mathrm{E}[Y] = \lambda \Gamma \{1 + 1/\gamma\} \tag{3.28}$$

and

$$\mathrm{Var}[Y] = \lambda^2 \left(\Gamma\{1 + 2/\gamma\} - \Gamma^2\{1 + 1/\gamma\}\right). \tag{3.29}$$

Let $\bar{X}_{Ak'}$ and $\bar{X}_{Bk'}$ denote the average response up to analysis k' for each treatment group. Again, as in the exponential case, the variability depends on the expected value. So, we will test the difference of their logged expected values and test, so $\mathrm{H}_0 : \log\lambda_A - \log\lambda_B = 0$ versus $\mathrm{H}_a : \log\lambda_A - \log\lambda_B = \delta$ or $\mathrm{H}_0 : \log\left(\frac{\lambda_A}{\lambda_B}\right) = 0$ versus $\mathrm{H}_a : \log\left(\frac{\lambda_A}{\lambda_B}\right) = \delta$. Our estimate at analysis k' is

thus $\log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right)$. Using the delta method and the central limit theorem it can be shown that,

$$\log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right) \sim \mathrm{N}\left(\log\left(\frac{\lambda_A}{\lambda_B}\right), \frac{2(\Gamma\{1+2/\gamma\} - \Gamma^2\{1+1/\gamma\})}{k' N_g \Gamma^2\{1+1/\gamma\}}\right). \tag{3.30}$$

The resulting standardized statistics are,

$$Z_{k'} = \log\left(\frac{\bar{X}_{Ak'}}{\bar{X}_{Bk'}}\right) \times \sqrt{k' N_g / 2\rho} \sim \mathrm{N}\left(\log\left(\frac{\lambda_A}{\lambda_B}\right)\sqrt{k' N_g / 2\rho}, 1\right) \tag{3.31}$$

where,

$$\rho = \frac{\Gamma\{1+2/\gamma\} - \Gamma^2\{1+1/\gamma\}}{\Gamma^2\{1+1/\gamma\}} \tag{3.32}$$

Similarly to section 3.1.2, the standardized test statistics follow the joint canonical distribution described earlier We thus have,

$$I_{k'} = \frac{k' N_g}{2\rho}, \tag{3.33}$$

which is the same as in the normal case with $\sigma^2 = \rho$. In a similar fashion as in the normal case, if we wish to test H_0 with significance level α and power $1-\beta$ at $\log(\lambda_A/\lambda_B) = \pm\delta$ our information level I at analysis k', for $k' = 1, \ldots, K$, is

$$I_{k'} = \frac{k' R M^2}{\delta^2}. \tag{3.34}$$

Equating (3.33) and (3.34),

$$N_g = \frac{2RM^2}{\delta^2}\rho \tag{3.35}$$

is the number of observed deaths needed for each treatment for each analysis. It is easy to see that the exponential distribution is a special case of the Weibull distribution with $\gamma = 1$. For illustration let's look at the example with $\gamma = 2$.

$$\mathrm{E}[Y] = \lambda\Gamma\{1.5\} \cong 0.886\lambda, \tag{3.36}$$

$$\mathrm{Var}[Y] = \lambda^2\left(\Gamma\{2\} - \Gamma^2\{1.5\}\right) \cong 0.2146\lambda^2, \tag{3.37}$$

and

$$\rho = \frac{\Gamma\{2\} - \Gamma^2\{1.5\}}{\Gamma^2\{1.5\}} \cong \frac{0.2146}{0.7850} = 0.2734. \tag{3.38}$$

From this, we have

$$N_g = 0.2734 \times \frac{2RM^2}{\delta^2} \tag{3.39}$$

Therefore, for any given α, β, and δ the required deaths, if the observation follow a Weibull distribution, is roughly 27% of that needed if the observation follow an exponential distribution.

3.1.5 Log Rank Test

From Jennison and Turnbull (1999), the sequence of score statistics $\{LR_1, \ldots, LR_K\}$ follows the joint canonical distribution defined in equation

(3.2) with,

$$LR_{k'} \sim \mathrm{N}\left(\theta I_{k'}, I_{k'}\right), \text{ for large } I_{k'}. \tag{3.40}$$

The information at analysis k' is just the estimated variance of $LR_{k'}$. We can standardize $LR_{k'}$ and the sequence $\{Z_1, \ldots, Z_K\}$ follows the joint canonical distribution in section 3.1, with

$$Z_{k'} = \frac{LR_{k'}}{\sqrt{\widehat{\mathrm{Var}}(LR_{k'})}} \sim \mathrm{N}\left(\theta\sqrt{\widehat{\mathrm{Var}}(LR_{k'})}, 1\right). \tag{3.41}$$

Again, under the null hypothesis it is expected that $r_{iA,k'} = r_{iB,k'}$, so that $\mathrm{Var}(LR_{k'}) = N_{k'}/4$. Hence, we can estimate the information needed at analysis k', for $k' = 1, \ldots, K$, to be,

$$I_{k'} \approx \frac{N_{k'}}{4}. \tag{3.42}$$

Note that this is the total information needed for both groups at analysis k'.

If we wish to test H_0 with significance level α and obtain a power of $1-\beta$ at $\theta = \pm\delta$ our required information at analysis k' for each treatment group is,

$$I_{k'} = \frac{k'RM^2}{\delta^2}. \tag{3.43}$$

Then the total required number of uncensored observations at analysis k' for each treatment group is

$$N_{k'} = \frac{4k'RM^2}{\delta^2} \tag{3.44}$$

To apply the results with equal group sizes we need equal information levels, so that the number of uncensored failures observed from one analysis to another is constant and equal to $2N_g$, where N_g is the required number of observation per treatment per analysis. Then $N_{k'} = 2k'N_g$ and,

$$N_g = \frac{2RM^2}{\delta^2}. \tag{3.45}$$

An interesting result: In a direct comparison to the results obtained for exponential responses in the two sample parametric hypothesis test, we see that we will need the same number of uncensored observation to achieve the same power in the the non-parametric log-rank test.

3.2 Nominal Significance Level

Here we will discuss two methods for repeated significance testing. In both cases the sequence of test statistics involved must follow those of the canonical joint distribution with equal incremental information levels at each analysis discussed in section 3.1. Essentially, the goal is to find nominal significance levels, $\alpha_i, i = 1, \ldots, K$, (or critical values) so that in the end the overall significance level is α. Here the nominal significance levels, the maximum number of observation and the number of interim analyses are chosen to achieve the overall significance level α (the probability of detecting a difference under H_0) and the power $1 - \beta$ (the probability of detecting a difference under H_a: $\theta = \pm\delta$). Thus given α, β and H_a we will be able to determine the sample size needed for each analysis and the total number of analysis, which in turn we will use to find the nominal significance levels needed at each analysis. Let K be the total number of analyses, and N_g be the number of uncensored observations on each treatment per analysis. In the models discussed below, K and N_g, must be chosen in advance. The number of deaths, $2N_g$, needed at each analysis is constant throughout. These are both definite restrictions on the flexibility of the group sequential approaches discussed. As both methods are relatively similar, I will discuss them in general, followed by a brief explanation of each individually, then give a brief comparison. An extensive one can be found in Jennison and Turnbull (1999). First, we will discuss the general theory behind the calculation of the nominal error levels.

3.2.1 Error Calculation

Suppose we have K interim analysis with a sequence of test statistic $\{Z_1, \ldots, Z_K\}$ that follow the canonical joint distribution defined in section 3.1 with information levels $\{I_1, \ldots, I_K\}$. Let $\tau_{k'} = I_{k'} - I_{k'-1}$. In the case of equal group sizes, $\tau_{k'}$ is constant.

At analysis k', let $(a_{k'}, b_{k'})$ be the stopping rule, such that if the test statistic $Z_{k'}$ is less than $a_{k'}$ or greater than $b_{k'}$ we stop the trial. The probability of exiting a trial at any specific point k' is,

$$\psi_{k'}(a_1, b_1, \ldots, a_{k'}, b_{k'}; \theta) = \mathrm{P}_{\mathrm{H}_0}(a_i < Z_i < b_i, i = 1, \ldots, k'-1; Z_{k'} > b_{k'}), \tag{3.46}$$

if the upper boundary is met at time k', and

$$\xi_{k'}(a_1, b_1, \ldots, a_{k'}, b_{k'}; \theta) = \mathrm{P}_{\mathrm{H}_0}(a_i < Z_i < b_i, i = 1, \ldots, k'-1; Z_{k'} < a_{k'}), \tag{3.47}$$

if the lower boundary is met at time k'. Recall from section 3.1, the distribution of $\{Z_1, \ldots, Z_K\}$ is multivariate normal. Thus, very tedious and difficult multiple integration over the multivariate normal distribution is required to calculate these probabilities. For values of K over 5, even the most sophisticated computers take a long time to calculate these numbers.

Wang and Tsiatis (1987) present a family of two-sided boundaries, $(a_{k'}, b_{k'})$, $k' = 1, \ldots, K$, indexed by a parameter κ. Different values of κ will produce different boundary shapes. Taking $\kappa = 0.5$ gives Pocock (1977), and $\kappa = 0$

gives O'Brien and Fleming (1979). These are special cases of boundaries with different κ. Other values of κ between 0 and 0.5 will produce intermediate boundary shapes. Given κ, the rejection of H_0 at analysis k' is determined by the boundaries,

$$a_{k'} = -c\,(k'/K)^{\kappa-0.5} \text{ and } b_{k'} = c\,(k'/K)^{\kappa-0.5}. \tag{3.48}$$

where c in some critical value. These critical values depend on K, α, and κ, or $c = C_{WT}(K, \alpha, \kappa)$. The null hypothesis is rejected at analysis k' if,

$$|Z_{k'}| > C_{WT}(K, \alpha, \kappa)(k'/K)^{\kappa-0.5}. \tag{3.49}$$

3.2.2 Pocock

The first method of repeated significance testing or groups sequential testing we will discuss is the method by Pocock (1977). This is a special example of the boundaries in (3.48) indexed at $\kappa = 0.5$ in (3.14). The boundaries for rejection of $Z_{k'}$ are simply, $\{-c, c\}$ for analysis k', $k' = 1, \ldots, K$. Here, $c = C_{WT}(K, \alpha, \kappa = 0.5) = C_p(K, \alpha)$. The critical value $C_p(K, \alpha)$ is a function of K, and α. However, in this case the boundaries remain constant throughout, and thus the nominal significance levels will be constant for each analysis. This is presented as a repeated significance test with constant nominal significance levels. Critical values in Table 5.3 are a replication of those found in Jennison and Turnbull (1999).

We reject H_0 at analysis k' if $|Z_{k'}| > C_p(K, \alpha)$, $k' = 1, \ldots, K$. $C_p(K, \alpha)$ is chosen so that,

$$P_{H_0}(|Z_1| > C_p(1, \alpha), \text{ or }, \ldots, \text{ or }, |Z_K| > C_p(K, \alpha)) = \alpha \tag{3.50}$$

The nominal significance level can be calculated as follows,

$$\alpha' = 2 \times P(Z > C_p(K, \alpha)) \tag{3.51}$$

where $Z \sim N(0, 1)$.

3.2.3 O'Brien and Fleming

The second method proposed by O'Brien and Fleming. This is a special example of the boundaries in (3.48) indexed at $\kappa = 0$. The boundaries for rejection of $Z_{k'}$ are now, $\{-c(k'/K)^{-0.5}, c(k'/K)^{-0.5}\}$ for analysis k', $k' = 1, \ldots, K$. Here, $c = C_{WT}(K, \alpha, \kappa = 0) = C_{OBF}(K, \alpha)$. The critical value $C_{OBF}(K, \alpha)$ is a function of K, and α. In this case however, the boundaries decrease throughout, and thus the nominal significance levels will be increasing in k'. In this test it is more difficult to reject H_0 in early stages and becomes easier later. An attractive feature with this method is that the significance level at the last analysis is very close to the overall significance level and thus resembles the fixed sample test. This minimizes the chance in which we would reject the null hypothesis in a fixed sample procedure but not in the group sequential test. This is a repeated significance test with increasing nominal significance levels. Critical values in Table 5.4 are a replication of those found in Jennison and

Turnbull (1999).

.We reject H_0 at analysis k' if $|Z_{k'}| > C_{OBF}(K,\alpha) \times \sqrt{K/k'}$, $k' = 1,,K$. The rejection rule becomes more liberal as k' increases, and thus easier to detect a difference in the end. $C_{OBF}(K,\alpha)$ is chosen so that so that the probability of rejecting the null hypothesis when it is true is α, or equivalently,

$$P_{H_0}\left(|Z_1| > C_{OBF}(1,\alpha)\sqrt{K/1}, \text{ or },\ldots, \text{ or },|Z_K| > C_{OBF}(K,\alpha)\sqrt{K/k'}\right) = \alpha \tag{3.52}$$

The nominal significance level can be calculated as follows,

$$\alpha'_{k'} = 2 \times P\left(Z > C_{OBF}(K,\alpha)\sqrt{K/k'}\right) \tag{3.53}$$

where $Z \sim N(0,1)$.

3.3 A Pure Sequential Design

Presented so far are group sequential methods where K, the number of interim analysis, must be fixed ahead of time. As well the number of observation per group, N_g, must be constant. Here we will explore a completely sequential design, where an interim analysis is performed after every event. Again we are testing the survival distribution between two treatments in a clinical trial. We assume the probability that a subject is assigned to treatment A is φ and 1-φ is the chance of being assigned to treatment B.

Gombay (2001) proposes two tests using U-statistics to compare two-treatments in a purely sequential design. At time $t = t_{(k)}$, we observe $k = k_A + k_B$ distinct events, $E_{(1)} < \ldots < E_{(k)}$. For each event there is an associated survival time, which is a measurement of the time the patient lasted in the study. Earlier we defined the survival times to be $X_j^{(y)} = min\left\{T_j^{(y)}, C_j^{(y)}\right\}$, $j = 1, \ldots,$ and $y = A, B$. If the subject realized the event and the exact survival time is observed then $X_j^{(y)} = T_j^{(y)}$, otherwise $X_j^{(y)} = C_j^{(y)}$. We denoted $C_j^{(y)}$ as the censored survival time. We indicate this by,

$$\Delta_j^{(y)} = \begin{cases} 1 & \text{if } X_j^{(y)} = C_j^{(y)}; \\ 0 & \text{if } X_j^{(y)} = T_i^{(y)} \end{cases} \tag{3.54}$$

where the distribution of Δ is identical for both groups under the null hypothesis. For the k observations at time $t_{(k)}$, we can thus summarize the observation as $\left\{Z_j^y\right\} = \left\{X_j^y, \Delta_j^y\right\}$. At time t_k we calculate,

$$U_k = \sum_{i=1}^{k_A} \sum_{j=1}^{k_B} h\left(X_i^{(A)}, X_j^{(B)}\right) \tag{3.55}$$

to compare the two treatments for $k \geq 2$. Gehan's statistic $h\left(X_i^{(A)}, X_j^{(B)}\right)$ is defined as,

$$h\left(X_i^{(A)}, X_j^{(B)}\right) = \begin{cases} 1 & \text{if } X_i^{(A)} > X_j^{(B)}, \Delta_j^{(B)} = 0; \\ 1 & \text{if } X_i^{(A)} = X_j^{(B)}, \Delta_i^{(A)} = 1, \Delta_j^{(B)} = 0; \\ 0 & \text{otherwise;} \\ -1 & \text{if } X_i^{(A)} < X_j^{(B)}, \Delta_i^{(A)} = 0; \\ -1 & \text{if } X_i^{(A)} = X_j^{(B)}, \Delta_i^{(A)} = 0, \Delta_j^{(B)} = 1. \end{cases} \tag{3.56}$$

A simpler version was offered by Mantel (1967), and illustrated in Miller (1981), where the combined sample from both groups was considered as $\{Z_1, \Delta_1\}$, $\ldots, \{Z_k, \Delta_k\}$ and calculated U_k with,

$$h\left(Z_m, Z_l\right) = \begin{cases} 1 & \text{if } Z_m > X_l, \Delta_l = 0; \\ 1 & \text{if } X_m = X_l, \Delta_m = 1, \Delta_l = 0; \\ 0 & \text{otherwise;} \\ -1 & \text{if } X_m < X_l, \Delta_m = 0; \\ -1 & \text{if } X_m = X_l, \Delta_m = 0, \Delta_l = 1. \end{cases} \tag{3.57}$$

Then, summing over the rows,

$$U_m^* = \sum_{l=1; l \neq m}^{k} h(Z_m, Z_l) \tag{3.58}$$

was calculated. The resulting test statistic U_k is calculated as the sum of the sums of the rows which correspond to observation from group A, and is calculated as

$$U_k = \sum_{m=1}^{k} U_m^* \pi_m, \tag{3.59}$$

where

$$\pi_j = \begin{cases} 1 & \text{if the j-th subject belongs to group A;} \\ 0 & \text{if the j-th subject belongs to group A.} \end{cases} \tag{3.60}$$

Under the null hypothesis it can be shown that

$$\widehat{\sigma}_k^2 = k^{-3} \sum_{i=1}^{k} \left[\sum_{j=1}^{k} h\left(Z_m, Z_l\right) \right]^2 \rightarrow \sigma^2 \text{ as } k \rightarrow \infty, \tag{3.61}$$

where $\widehat{\sigma}_k^2$ is the estimated variance of the test statistic $k^{-3/2}U_k$ at analysis k.

Two tests are proposed in Gombay (2001) under the criteria mentioned above, truncated after n observations. This truncation point is the maximum number of patients allowed to enter the study.

Test 1. For $k = 2, 3, \ldots, n$, we calculate the test statistics,

$$Z_k = \frac{k^{-3/2}|U_k|}{\widehat{\sigma}_k \sqrt{\varphi(1-\varphi)}}. \tag{3.62}$$

The test is rejected the first time the test statistic Z_k exceeds the critical value $CV_1(\alpha, n)$, otherwise do not reject H_0. These values can be found in the appendix in Table 5.5, and are described in Gombay (2001). **Test 1** is similar to fully sequential Pocock boundaries. This test offers more conservative results and is not expected to perform as well as **Test 2**. **Test 1** will not be considered in this thesis.

Test 2. For $k = 2, 3, \ldots, n$, we calculate the test statistics,

$$Z_k = \frac{k^{-1} n^{-1/2} |U_k|}{\widehat{\sigma}_k \sqrt{\varphi(1-\varphi)}}. \tag{3.63}$$

Stop and reject H_0 the first time Z_k exceeds the critical value $CV_2(\alpha, n)$, otherwise do not reject. These values can be found in the appendix in Table 5.5, and are described in Gombay (2001).

Chapter 4

Simulation and Results

4.1 The Simulation

Regardless of whether a statistical model requires knowledge of the distribution of certain random variables, we cannot simulate its performance without somehow generating these random variables. To generate these random numbers we must assume some underlying distribution. Upon entry each subject was randomly allocated to one of two treatment groups with the probability of assignment to group A being φ. Here we chose $\varphi = 0.5$ so that we have equal allocation to each group. For the sequence of independent arrivals $U_j^{(y)}, j = 1, \ldots, n$, and $y = A, B$, we assumed uniform entry over a 10 year period. If U is said to be uniform in (a,b), $U \sim \mathrm{U}(a, b)$ then

$$f_U(t) = \frac{1}{b-a} \ , \ a \leq t \leq b \ , \tag{4.1}$$

with $\mathrm{E}[U] = \frac{a+b}{2}$ and $\mathrm{Var}[U] = \frac{(b-a)^2}{12}$. $U_j^{(y)} \sim \mathrm{U}(0, 10)$. Most programming software has built in functions for generating some type of uniform random numbers. The key to generating any random phenomena is to generate a U(0,1) random number which we will call u. For the simulation of entry time

we generate u, then $U_j^{(y)} = 10 \times u$.

We assumed the survival times $X_j^{(y)} = \min\left\{T_j^{(y)}, C_j^{(y)}\right\}$, $j = 1, \ldots, n$, and $y = A, B$, for the independent individuals are exponentially distributed. For an exponential random variable X with $\mathrm{E}[X] = 1/\lambda$, the cumulative distribution function is,

$$F_X(t) = 1 - e^{-\lambda t} \in [0, 1] \ , \ t \geq 0. \tag{4.2}$$

We generate a U(0,1) random number u. Set $u = 1 - e^{-\lambda t}$, we solve for t so that,

$$t = -\frac{1}{\lambda}\log(1 - u). \tag{4.3}$$

A truncation point n with $\omega_1 = 0$ (no chance of loss) is calculated using $SPlus^{®}$. The final truncation point is calculated with the probability of type I censoring chosen to be $\omega_1 = 0.05$ as mentioned before.

4.1.1 The Algorithm

The simulation involved simulating the trial for each individual n subjects. For subject k, $k = 1, \ldots, n$: (Generate the vector $[U, X, \pi, \delta, E]$ for each subject)

1. Generate $u_{1_k} \sim \mathrm{U}(0, 1)$: then $U_k = 10u_{1_k}$;

2. Generate $u_{2_k} \sim \mathrm{U}(0, 1)$: if $u_{2_k} < 0.5$ then $\pi_k = 1$ (allocate the patient to treatment A), otherwise $\pi_k = 0$ (allocate the patient to treatment B) ;

3. Generate $u_{3_k} \sim \mathrm{U}(0,1)$: if the patient is in treatment A then the survival time $X_k = -\frac{1}{\lambda_A}\log(1-u_{3_k})$, otherwise the patient is in treatment B and $X_k = -\frac{1}{\lambda_B}\log(1-u_{3_k})$;

4. Event time $E_k = U_k + X_k$;

5. Generate $u_{4_k} \sim \mathrm{U}(0,1)$: if $u_{4_k} < \omega_1$ (ω_1 = the probability of type I censoring = 0.05) or $E_k > 10$, then $\Delta_k = 1$, otherwise $\Delta_k = 0$;

6. Sort the vector by event time. We now have n distinct ordered event times $E_{(1)} < \ldots < E_{(n)}$.

7. Calculate the $n \times n$ matrix using the kernel function in (3.57). This is called the Hmatrix;

8. Calculate the UMatrix and the UMatrixAll;

9. Calculate the UStat_k (3.59) for $k = 1, \ldots, n$;

10. Calculate $\hat{\sigma}_k^2$ (3.61) for $k = 1, \ldots, n$;

11. Calculate the test statistic for **Test 2**, Z_k (3.63) for $k = 1, \ldots, n$;

12. Compare Z_k to the $CV_*()$;

13. Terminate the simulation the first time $Z_k > CV_*()$ and reject H_0 (increment the counter $b() = b() + 1$, k is the stopping time (increment the counter $Obs() = Obs() + k$);

14. If no Z_k reached $CV_*()$, H_0 is not rejected ($b() = b()$, $Obs() = Obs() + n$);

4.2 Case 1: Comparison using U-Statistics

We will denote the group sequential method by Pocock (1977) as P, the groups sequential approach by O'Brien and Fleming (1979) as OBF, and the fully sequential procedure by Gombay (2001) as **Test 2**. The first analysis we consider is comparing the fully sequential method to group sequential methods using the U-statistic with the kernel defined in (3.56). Here, we fix the truncation point, n, the significance level, α, and the number of interim analysis under the group sequential approach, K. Here we will ignore type II censoring so that all patients will be followed until their time of death. Type I censoring will still be considered using the Gehan statistic (3.56) with $\omega_1 = 0.05$. Under these conditions data is simulated using the algorithm already explained. For the fully sequential method the algorithm explains the stopping criteria. Using the U-statistic under a group sequential approach, with K interim analysis, we need to determine the appropriate boundary values at each analysis k', $k' = 1, \ldots, K$. Using the same test statistic in **Test 2** (3.63) we must determine the rejection rules for P, and OBF. In the group sequential approach we will only look at the data K times. With a truncation point of n, we analyze the data at times $k = (nk'/K)$ for $k' = 1, \ldots, K$. For a groups sequential approach we analyze the data at k, $k = n/K, 2n/K, 3n/k, \ldots, n(K-1)/K, n$. Thus for fixed n and K, the time if analysis is a function of k'. The test statistic Z_k (3.63) for **Test 2** can be written in the form,

$$Z_k = (K/nk')\frac{n^{-1/2}|U_k|}{\widehat{\sigma}_k\sqrt{\varphi(1-\varphi)}} = \frac{S_k}{\sqrt{n}}, \tag{4.4}$$

for some S_k.

In the group sequential notation $Z_{k'}$ is the same as Z_k, only that we have restricted values of k at, $k = n/K, 2n/K, 3n/k, \ldots, n(K-1)/K, n$. For example if we have K=5 and n=100, in the group sequential notation, $Z_{k=20}$ is equivalent to $Z_{k'=1}$. The group sequential rejection rules described in section 3.2.2 and 3.2.3 are based on test statistics Z_k in the form, $Z_k = S_k/\sqrt{k}$, Jennison and Turnbull (1999).

To make a correct comparison we must multiply the test statistic in (4.4) by $\sqrt{n/k}$. Therefore, from section (3.2.2), for P, we reject H_0 at analysis k if $|Z_k|\sqrt{n/k} > C_P(K,\alpha)$ or when $|Z_k| > C_P(K,\alpha)\sqrt{k/n} = C'_P(k,K,\alpha)$. Critical values $C_P(K,\alpha)$ are found in table 5.3. For example, if we take $K = 5$ and $\alpha = 0.05$, we have $C_P(5, 0.05) = 2.413$. We will reject the null hypothesis at analysis 1 if $|Z_{k=n/5}| = |Z_{k'=1}| > 2.413 \times \sqrt{1/5} = 1.079$. Similarly, $C'_P(2(n/5), 5, 0.05) = 1.525$, $C'_P(3(n/5), 5, 0.05) = 1.869$, $C'_P(4(n/5), 5, 0.05) = 2.158$, and $C'_P(5(n/5), 5, 0.05) = C_P(5, 0.05) = 2,413$.

Similarly, from section (3.2.3), for OBF, we reject H_0 at analysis k if $|Z_k|\sqrt{n/k} > C_{OBF}(K,\alpha)\sqrt{K/k'}$ or simply when $|Z_k| > C_{OBF}(K,\alpha)$. Critical values $C_{OBF}(K,\alpha)$ are found in table 5.4.

4.2.1 Results for Comparison 1

The results in tables 5.8 - 5.11 are based on simulation using the algorithm in section 2.2. Results are obtained for values of $n = 100$, 150, 200 250, $K = 2$,

5, 10, and $\lambda_h = 1.0$, 1.5 and 2.0. The results are from 10000 simulated trials. For the group sequential methods, the algorithm remains unchanged, however we only compare Z_k to the critical value at times nk'/K, for $k' = 1, \ldots, K$. Also remember, the results for **Test 2** do not depend on K so the results remain constant as K changes. The tables show the average stopping time of the study in terms of number of events, and power of the test. The average stopping time is the average number of patients to leave the study (by either dying or being lost) before a decision can be made. If the null hypothesis is never rejected then the stopping time will be the truncation point n. The truncation point is the essentially the stopping time in a fixed sample setting or $K = 1$.

For the group sequential approaches we have discussed the requirement of equal information levels (or equal number of uncensored observations). Jennison and Turnbull (1999) discuss the effects of slight deviation from equal group sizes. It is shown that minor deviations will have little to no effect on the power and significance levels of the test. For this application we set the truncation point at n with probability of type I censoring is $\omega_1 = 0.05$. Thus we'd expect to see 0.95n uncensored events (or deaths). In any case for K analysis, we will inspect the data at time nk'/K, $k' = 1, \ldots, K$. This will achieve equal group sizes in terms of the number of patients to leave the study. This will not however guarantee equal group sizes in terms of the number of required deaths (or uncensored observations) at each analysis. However, based on the constant probability of a patient being lost to follow up, we will expect to see the same number of deaths for each analysis. However, due to randomness this will vary, but only slightly, having little to no effect on the

results. For example if we take $n = 200$ and $K = 5$, we will inspect the data at $N = 40, 80, 120, 160$ and 200. If the probability of type I error is $\alpha = 0.05$ we will expect to see 190 uncensored observations. This would require inspections after 38, 76, 115, 152 and 190 deaths. For example, due to randomness, we may inspect after 37, 77, 111, 154, 189 deaths. These slight deviations will have very little effect on the results.

The other result we are concerned with is the power of the test. For values of λ_h other than 1 we would expect to reject the null hypothesis, $H_0 : \lambda_h = 1$. The power of the test measures the probability that the test would detect a difference when one exists. For $\lambda_h = 1$ we would expect the power of the test to equal the significance level α.

The results illustrate no obvious differences for different values of n. We can definitely see an improved power for all methods, but for comparisons sake the difference in power between the methods is relatively constant for all values of n.

For $K = 2$, **Test 2** has lower average stopping time than the two group sequential methods with OBF being the slowest to reject H_0. However, the power for OBF is the best with the power for **Test 2** and P being very comparable. The difference in power decreases for larger differences in treatment means.

As K increases to 5, we can see a decrease in power for the group sequential methods. The power for P drops off the most falling quite a bit below

Test 2. The decrease in power for OBF isn't as dramatic with this method remaining the most powerful. The stopping times decrease as well for P, now the fastest to reject H_0, followed by **Test 2** then OBF.

For $K = 10$, the changes are similar and the results are similar to K=5. Again, the power and average stopping time decrease for the group sequential methods. OBF remains the most powerful with P the fastest to stop. **Test 2** performs in the middle for both categories.

For $\lambda_h = 1$ we have no difference in the treatments; this is the null hypothesis. With a significance level of $\alpha = 0.05$, we would expect to reject the null hypothesis 5% of the time when it is in fact true. From Tables 5.8 - 5.11 we can see all the results were close to 0.05. The results of **Test 2** were generally slightly less than 0.05, while the methods P and OBF were consistently on the high side. This would indicate a more conservative result from **Test 2**, which may have a slight effect of under estimating the power for other values of λ_h. In any case, the smaller α the better, and we consistently see a better result from **Test 2**.

In the end we see that the fully sequential approach is very comparable, in both power and stopping time and has a smaller type I error. **Test 2** generally performs better in power than P and stops quicker than OBF. As well, the fully sequential model also offers complete flexibility to do analysis whenever is convenient.

The graphs of stopping boundaries in the appendix illustrate this for the

cases when $K = 5$ and, 10 and $\alpha = 0.05$. We see the boundaries for **Test 2** using fully sequential boundaries, as well as the group sequential boundaries for P and OBF. For reference the fixed sample boundary is included. Note, the thin lines for the group sequential or fixed sample boundaries are imaginary. Rejection can only occur at the dots. For the fully sequential procedure, rejection can occur at anytime, hence the thick boundary line.

For $K = 10$: We can see the early stopping conditions for P are much more relaxed than for **Test 2**. In the end, however, the boundaries for P are much more strict making it hard to detect a difference. This creates a confusing situation where the null hypothesis could be rejected if a fixed sample test were performed, but P would detect no difference. If P does not reject the null hypothesis (when it is false) early on, it may difficult to ever reject. This reinforces the results where we found P to be slightly faster to stop than **Test 2**, but much less powerful. Essentially, when P stops, it stops early, otherwise doesn't stop at all, and no difference may be detected in the end. Similar results are seen for $K = 5$.

OBF offers a slightly improved approach to P in terms of its ability to detect a difference. The boundaries for OBF are always slightly less strict than for **Test 2**, making it obviously easier to detect a difference at time of analysis. The key is that this detection can only be at one of the interim analysis, making the stopping time for OBF to be much longer than **Test 2**. We found that OBF was slightly more powerful than **Test 2** (because of the easier stopping rules), but takes longer on average to detect a difference (because of the restricted analyses times). Similar results are seen for $K = 5$.

Although it is notable to mention that as K increases the boundaries for OBF approach those of **Test 2**.

4.3 Comparison 2: U-Statistic vs. Sequential Log-Rank

This application will compare the results between two different models, the U-statistics versus the log-rank test. We will compare fully sequential U-statistics against group sequential log-rank statistics. We should note we are comparing the U-statistics versus the best possible result in the log-rank test. These results are more of an indication of which model is the best rather than a comparison of the group sequential versus fully sequential procedures.

We are again investigating the power and stopping time of sequential methods. For the group sequential methods we used the canonical joint distribution of K sequential log-rank statistics to determine the number of deaths required to achieve some power 1 - β under the alternative hypothesis that $\theta = \pm\delta$ at significance level α. These values are also dependent on the nominal significance levels chosen for the K interim analyses. Once the number of deaths needed was found we used the $S + SeqTrial$ module in $SPlus^{®}$ to calculate the corresponding truncation point needed given the length of the trial and distributions of the entry and survival times. Here we will assume a 10 year study time with entry time being uniformly distributed throughout. To calculate the truncation point the function $SeqPHSamplesize$ in $SPlus^{®}$ assumes uniform entry and exponential survival times. This is the truncation point that is used in the simulation.

Under the group sequential log-rank test with censoring we can now calculate the truncation point for the combinations {α, δ, β, K}. We performed 10000 simulation of the pure sequential method for different levels of the same

$\{\lambda_A, \delta, \mathrm{K}\}$ at the corresponding truncation point required to achieve the power 1- β. The simulation gave us the power of the test and the average stopping.

Suppose we are testing the null hypothesis $\mathrm{H}_0 : \theta = \log\left(\frac{\lambda_A}{\lambda_B}\right) = 0$ versus the alternative $\mathrm{H}_a : \theta = \log\left(\frac{\lambda_A}{\lambda_B}\right) = \pm\delta$. In sequential testing, whether it be group sequential or pure sequential, we reject the null hypothesis the first time the test statistic exceeds the critical value of the test at some significance level, α. For the tests discussed, critical values are calculated for different significance levels α and some truncation points n. The truncation point is the total number of observations allowed in the study.

We wish to compare the results obtained from the group sequential log-rank test using Pocock (1977) and/or O'Brien and Flemings (1979) significance level approaches to Gombay's **Test 2** in the pure sequential method using U-statistics. The goal is to determine the relative performance of the U-statistics to that of the log-rank statistics. To do this we must control the characteristics of each analysis. We are mostly interested in comparing the power of the tests as well as the average stopping time. The power of the test measures the ability to detect a difference in treatments when one exists, and the stopping time is the time to rejection of the null hypothesis. If the null hypothesis is never rejected, the stopping time is just the length of the study.

Power and stopping time depend on many factors. In the group sequential procedures we obtain these results for different levels of α (the significance level), at some δ (the magnitude of the treatment difference under the alternative hypothesis) and K (the number of interim analysis). In the pure

sequential procedure the power and average stopping time depend on α, δ and the truncation point.

For a comparison to make sense we must make sure the characteristics of the two procedures are equivalent. The simulation for the fully sequential test, **Test 2** requires that we know α, δ and the truncation point. The first two, α and δ, are easy enough to fix. However, since the results from P, OBF only report the number of deaths needed, we must find a way to calculate the truncation point to be used in the fully sequential approach. We will first determine the required number of deaths from the group sequential procedures at different combinations of α, δ, K and the power, 1-β. From this, considering the length of the study, the distribution of the entry times, the distribution of the survival times, as well as some probability of loss we can obtain the expected number of patients needed to satisfy the information requirement. These number of patients will be the desired truncation point used in the fully sequential simulation.

The simulation will then be performed at the same significance level α and the same value δ used to calculate the truncation point. In the group sequential procedures we know what the power of the test and the average stopping time will be given these K, α, δ and n. The simulation will estimate the power of the test and the average stopping time given α, δ and the truncation point. We wish to compare the results obtained through simulation of the pure sequential procedure versus those expected using the group sequential procedure.

4.3.1 The Truncation Point, n

The truncation point is the total number of patients in the study. To obtain n we must first determine the number of deaths required to achieve the power $1-\beta$ at some alternative $\pm\delta$.

In clinical trials we have mentioned that censoring is very common and cannot be avoided. We should not ignore these observations. Given that censoring will occur we must determine the number of patients needed to obtain the required number of events. The number of events is determined by the study design with significance level α to achieve a certain power 1-β at $\pm\delta$. To do this we must determine a probability model for the expected number of censored observations. The truncation point will just be the number of expected censored observations plus the required number of deaths. Recall, we have two types of censoring. There is a possibility a patient is lost during the study. In which case we have to assign a value to this probability. We will denote the probability of losing a patient as $P(lost) = \omega_1$. There are no obvious choices for ω_1. The other type of censoring is when the patient survives the study. This probability, which we will call $P(survived) = \omega_2$, depends on when the patient entered the study, the patients survival time and the chosen length of the study. The length of the study will be referred to as the maximum time frame allowed before a conclusion must be made. In this thesis our results will be based on a study that lasts 10 years. We denote a patients entry time to be $U_j^{(y)}$ for $j = 1, 2, \ldots$, and $y = A, B$, and we will assume uniform entry for the time of the study so that $U_j^{(y)} \sim \mathrm{U}(0, 10)$. In non-parametric models no assumption of survival time is made. However, for simulation purposes we need to generate random survival times based on some probability distribution. For

comparison's sake, we will assume the patients survival time is exponential. Recall, we are testing the null hypothesis $H_0 : \theta = \log\left(\frac{\lambda_A}{\lambda_B}\right) = 0$ versus the alternative $H_a : \theta = \log\left(\frac{\lambda_A}{\lambda_B}\right) = \pm\delta$. Let the survival times from treatment A, $X_j^{(A)}$, be exponential with λ_A and the survival times from treatment B, $X_j^{(B)}$, be exponential with λ_B, for $j = 1, 2, \ldots$. We then have the probability that a patient in group A survives to be $P(U + X_A > 10)$, and the probability that a patient in group B survives to be $P(U + X_B > 10)$. Given the probability of assignment to treatment A is φ, the overall probability that any given patient survives the study is

$$\omega_2 = \varphi P(U + X_A > 10) + (1 - \varphi)P(U + X_B > 10). \tag{4.5}$$

If we assume that patients are equally allocated to the treatments, then $\omega_2 = 0.5\{P(U + X_A > 10) + P(U + X_B > 10)\}$. These probabilities are not calculated. However the result $N/(1 - \omega_2)$ is obtained using sequential testing module $S + SeqTrial$ in $SPlus^{®}$. Recall, N is the total number of deaths (uncensored failures) needed to achieve a power of 1 - β at δ. A function in $SeqPHSampleSize$ calls us to provide the number of required deaths, the length of the study, the median survival time of the control treatment (treatment A), and the hazard ratio under the alternative hypothesis (δ). The function calculates the estimated number of patients required in the study so that we will observe N uncensored deaths only taking in account only the second type of censoring. The output received is thus $N/(1 - \omega_2)$. The function assumes uniform entry and exponential survival times, this is why our simulation will do the same.

Now, based on this result and ω_1 we can calculate the expected number of patients needed in a 10 year study so that on average we will observe the N required deaths. We denote n as the truncation point and N as the number of deaths. Thus,

$$n = \frac{N}{(1-\omega_1)(1-\omega_2)}. \tag{4.6}$$

In the simulation study, based on ω_1, we will check for type I censoring first. Then if there is no type 1 censoring, we check for type II censoring.

4.3.2 Sequential Log Rank Test

For the group sequential procedures discussed we will use the canonical joint distribution of the log-rank scores. Using the results obtained from the sequential log-rank test, we showed how to calculate the required number of deaths for each interim analysis for each group to be,

$$N_g = \frac{2RM^2}{\delta^2}. \tag{4.7}$$

where, R and M are defined in section 3.1.1. The results of P, and OBF require equally spaced information levels. Thus, for K interim analysis we require a total of,

$$N = 2KN_g = 2K\frac{2RM^2}{\delta^2} \tag{4.8}$$

deaths to achieve the power 1- β at $\pm\delta$. Thus to satisfy the power requirement at $\pm\delta$, given the distribution of entry and survival times, as well as the probability of losing a patient we need a maximum sample size,

$$n = \frac{4KRM^2}{(1-\omega_1)(1-\omega_2)\delta^2}. \tag{4.9}$$

Values of N_g can be calculated for various levels of α, 1 - β, δ for each group sequential procedure using the formulas above. Examples were calculated and confirmed in $SPlus^{®}$.

4.3.3 Calculating the Truncation Point, An Example

Again we are testing $H_0 : \theta = \log(\lambda_A/\lambda_B) = 0$ versus $H_0 : \theta = \log(\lambda_A/\lambda_B) = \pm\delta$. At significance level $\alpha = 0.05$, for $K = 5$ analysis we want to determine the required sample size to achieve a power of 1-$\beta = 0.9$ at $\lambda_A/\lambda_B = 1/1.75$. Using Pocock's significant level approach we get $R_P(0.05, 0.9, 5) = 1.207$. We want to calculate $N_g = \frac{2RM^2}{\delta^2}$. From the results in Chapter 3 using the joint canonical distribution we found that the required number of events is the same for the two sample test with exponential responses as it is for the log-rank statistics. Our results showed that by using the canonical joint distribution we can calculate the required sample for any sequence of test statistics by using the two sample test with normal means. In any case, we calculate a value of $N_g = 16.2$, the number of deaths required per analysis per treatment arm.

Using S-Plus

In *SPlus*® we can calculate the required number events by using either a hazards ratio model with exponential responses (because of the equivalence of the statistics under the canonical joint distribution) or a normal means model. Here I will illustrate the use of the joint canonical distribution for the hazard ratio model for use in the normal means model.

The hazard ratio model for exponential responses in *SPlus*® is set up with the hazard ratio defined as $\lambda_h = \frac{\lambda_A}{\lambda_B}$, where λ_A is the hazard rate of the control group. Consider the alternative hypothesis $\delta = \pm\log(\lambda_h) = \log(1/1.75)$. *SPlus*® requires we enter δ, K and the boundary shape with parameter κ (section 3.2.1). Using this function *SPlus*® gives us the same results as we calculated using the canonical joint distribution for the hazard ratio model with exponential responses. Since we've shown the results for the hazard ratio model are the same as for the log-rank statistics we have our result. *SPlus* gives the value of $N_g = 16.2$.

However, we can show that our result can also be obtained using the normal means model. The normal means model is testing the difference between the means of two populations with normal responses. The normal means function requires we enter the hypothesized mean difference, the alternative mean difference, the standard deviations of the two populations, the number of analysis and the boundary shape. Using the log-rank test we have to use the joint canonical distribution theory to apply these models to the normal response model. As seen in chapter 3, the log-rank test, based on the canonical

joint distribution, is equivalent to a normal means test with the alternative $\pm\delta = \pm(\log(\lambda_A) - \log(\lambda_B)) = \pm\log(0.57142) = \pm 0.5596$ and $\sigma^2 = 1$. If a sequence of test statistics can be formulated to fit the canonical joint distribution then we can use the normal means model for the sequence of standardized test statistics. From now on we will focus on the sequential log-rank test.

Now that we have established the process of calculating the required number of deaths, let's return to our example and calculate the truncation point. In order to compare the results from the sequential log-rank test using P or OBF to the fully sequential **Test 2** we must determine the truncation point. If no censoring can occur then the number of required deaths, N, is the truncation point, n. However based on some probability of censoring (not 0) we will require a truncation point larger than the number of deaths.

We will assume the control treatment has a hazard rate of $\lambda_A = 1.5$, which equates to a median response of 1.0395. Note that for any exponential random variable X with expected value $1/\lambda$, the cumulative distribution function is

$$F_X(t) = \mathrm{P}(X \leq T) = \int_0^t \lambda e^{-\lambda s} ds = 1 - e^{-\lambda t}\ ,\ t \geq 0, \tag{4.10}$$

where the median, t is such that $\mathrm{P}(X < t) = 0.5$, or

$$1 - e^{-\lambda t} = 0.5 \ \Rightarrow\ t = -\frac{1}{\lambda}\log(0.5). \tag{4.11}$$

Assume the study will last 10 years and entry is $\mathrm{U}(0, 10)$. At $K = 5$, $\alpha = 0.05$, $1 - \beta = 0.9$ and $\delta = \log(\lambda_h) = \log(1/1.75)$ we need a total

of $N_g = 16.2$ or rounding up to the nearest integer, $N_g = 17$ events per treatment per analysis using significance level approach P. In total, we will need $N = 2KN_g = 10N_g = 170$ deaths to meet the power requirement at $\delta = \log(\lambda_h)$. Based on the probability of type 2 censoring, $\mathrm{P}(survival) = \omega_2$, $S - plus^{®}$ calculates the required sample size on average to obtain 170 events in 10 years to be $N/(1 - \omega_2) = 213.4$. This value is the required sample size needed before accounting for the probability of loss. Taking this into account and assuming the chance of type I censoring or, $\mathrm{P}(loss)=\omega_1 = 0.05$, our truncation point is thus,

$$n = \frac{170}{(1-\omega_1)(1-\omega_2)} = \frac{213.4}{(1-\omega_1)} = \frac{213.4}{0.95} = 223.68. \tag{4.12}$$

Thus, for comparisons in the simulation study, we would set $\alpha = 0.05$ (this will determine our critical values), $\lambda_A = 1.5$, $\lambda_B = 2.625$ and simulate the study for $n = 224$ patients. The simulation will produce estimates for the power of the test which will be compared directly to 0.9, the power for P. As well, we will simulate the average number of deaths before stopping.

The results for the required information levels and thus the number of deaths per treatment per analysis for the sequential log-rank test can easily be obtained using the results from (3.45). Then the truncation point is calculated after taking into account the censoring probabilities. Using the $S + SeqTrial$ module in $SPlus^{®}$ we can determine the average stopping time for the group sequential methods.

4.3.4 Results for Comparison 2

The first comparison is done strictly with respect to the results from the group sequential methods. All the results are completely relative to those of P and OBF using the log-rank statistic. That is, as we will see later, as δ increases the power of the pure sequential method deceases with the power held constant for the group sequential procedures. This is due to the fact we are fixing 1-β for the group sequential procedures. By fixing the power we will need different truncation points for different values of λ_A and λ_h. That does not mean that the pure sequential procedure gets weaker for larger treatment differences, only that its power does not increase as much as P or OBF.

We fixed the power, 1 - β, in the group sequential methods and determined the required truncation point to meet this power at some alternative $\pm\delta$. Then we simulated the pure sequential method using this truncation point to see how well it performs relative to P and OBF. Define λ_h as the hazard ratio and λ_A as the hazard rate for the control group. Under the alternative hypothesis $\log(\lambda_h) = \pm\delta$. The significance level, α is fixed at 0.05 throughout. For the combinations $\{\beta, K, \delta, \lambda_A\}$ a truncation point was calculated for each of P and OBF, where $1 - \beta = 0.8, 0.9$, $K = 2, 5, 10$, $\delta = 1/1.75, 1/2.00$, and $\lambda_A = 1, 1.5, 2$. Note, $\lambda_h = \frac{\lambda_A}{\lambda_B}$. Using the algorithm already explained we simulated the results for 10^5 replications. Tables 5.12 to 5.15 show the results at $\alpha = 0.05$.

The first two tables show the results when compared to the significance level approach used by P where the nominal significance levels are constant for each analysis while keeping the overall significance level α constant at 0.05.

The second two tables are the results when compared to OBF where initially it is difficult to detect a difference but assigns more chance in the later stages with the last analysis closely resembling a fixed sample test.

As the number of interim analyses increase, the performance of **Test 2** increases relative to the others. For $K = 2$ analysis, there is definitely inferiority problem, as we see the power around 0.65 where it is 0.8 for both group sequential methods. However, as K increases we can see a definite improvement where at $K = 10$ the powers are very close in comparison to P, although still far off from OBF. This is expected for any K since in the end we know OBF will have a higher nominal significance level than P.

As the magnitude of δ increases we see a slight decrease in power relative to the power in the group sequential procedures. Of course the power will increase as δ increases, but this analysis shows that the increase isn't as much as the increase in the group sequential methods. Holding δ constant, we can see that as λ_A gets larger our relative power increases as well.

One more observation is that the pure sequential method performs much better in comparison to P than it does compared to OBF. Regardless of the chosen power there doesn't appear to be any relative impact. This analysis was more of an experiment to observe the relative effects of some factors on the performance of the tests. Here, no comparison was done for the stopping time. From the tables we can see that **Test 2** test improves with respect to P and OBF as K increases.

Now we look at the comparison from a different perspective. Here we fix the truncation points and compare results. The following tables show a direct comparison of power and average stopping time for fixed truncation points. In this application consider the average stopping time to to be the average number of observed (uncensored) events before stopping. That is, we will compare the average number of deaths until we can stop the study. For the chosen truncation points of $n = 100, 150, 200, 250$, we simulated 10^5 results for $\lambda_h = 0.57143$ (or $1/1.75$) , 0.5 (or $1/2$) , at $\lambda_A = 1, 1.5, 2$. Here, we fix $K = 5$ for the group sequential methods. Again, for $\alpha = 0.05$, we calculated the associated powers and average stopping times for P and OBF using $SPlus$®. Now, unlike the earlier analysis, we must now determine the number of deaths we will expect to see given a certain truncation point. I am now fixing the truncation point and by $SeqPHSampleSize$ in $SPlus$® and ω_1 we can determine the associated number of events we will expect to see after censoring and use this number to get the power and average number of deaths. This is done in $SPlus$®. For example, for a sample size of $n = 100$ in Table 5.16, I will expect to see 83.22 deaths given uniform entry, hazard rate λ_h with λ_A, in a 10 year period. I will use 83.22 to obtain the result for P and OBF in $SPlus$®. For **Test 2** I will use the truncation point of $n = 100$. Table 5.20 shows the expected number of deaths versus the observed number of death after 10^5 simulations for $n = 100, 150, 200, 250$.

In any case, this analysis offers a more direct approach to compare the methods at different levels of δ. Here we compare the power and the average number of deaths. Tables 5.16 to 5.19 show the results. As δ increases we can see a close comparison of power between the **Test 2** and the group sequential

log-rank test. For $n = 150$ all three methods perform very well for large values of δ. This is done for $K = 5$, where the pure sequential method performs competitively at $n = 150$, 200 and 250. At $\delta = 1/2.5$ (0.4 in the normal means model with $\sigma^2 = 1$), we see little difference between the powers in either of the methods. The results show that for large treatment differences there is little difference between the powers of the tests.

Tables 5.16 - 5.19 also show the results of the average number of death before stopping. The truncation points are set at $n = 100$, 150, 200, 250. The tables show the expected number of deaths under the fixed sample test, where we would wait until the end of the ten years to analyze the data, compared to the number deaths we'd expect to observe under the three sequential procedures. For $K = 5$ it appears that method P performs the best then OBF, followed by **Test 2**. For small values of λ_h, **Test 2** performs relatively well with the stopping time only being slightly longer. As λ_h increases we see the groups sequential log-rank test performs much better than **Test 2**. Regardless we can still see a definite improvement in the average number of deaths over a fixed sample test.

To conclude, the method using the U-statistics did not perform as well compared to those of log-rank statistics. For large differences under the alternative hypothesis we see little difference in power. In terms of average number of deaths, the U-statistic model tends to take longer to detect a difference. These results are expected, and are not a comparison of fully sequential tests versus group sequential tests as much as they area comparison of U-statistics versus log-rank statistics (the best case scenario).

4.4 Final Conclusions

This thesis commented little on the comparison of methods P to OBF as there is an extensive comparison in Jennison and Turnbull (1999).

Comparison 1 is of primary interest showing a direct comparison of a fully sequential procedure to a group sequential approach under the same model using U-statistics. Here we see that the fully sequential model has its benefits with respect to P, and OBF. We saw the fully sequential method generally is more powerful than P and stops quicker than OBF.

Comparison 2 showed a comparison of two different models. We found the best possible case, the group sequential log-rank test, performed better than the fully sequential **Test 2** using the U-statistics. This is not surprising considering the results from the canonical joint distribution of the log-rank statistics. We found the group sequential log-rank test perform similar to a two sample test with exponential responses. So in the end we were essentially comparing a Gombay's (2001) **Test 2** to a parametric test. Given we simulated exponential responses we would expect a parametric test to perform better.

In the end, all methods have their advantages. Using U-statistics (Comparison 1), Pocock (1977) gives the best opportunity of early stopping, sacrificing power in the end. As time progresses detection becomes more difficult. O'Brien and Fleming (1979) has the best chance of detecting a difference, if one exists, but has the longest stopping time. Gombay's (2001) fully sequential **Test 2** offers some middle ground. **Test 2** is more powerful than Pocock

(1977) and stops quicker than O'Brien and Fleming (1979).

The pure sequential method has other benefits. We benefit from the fact that we can look at the data whenever we want and however many times we want. Regardless of it being a pure sequential method, if we only want to look at the data K times, we can. The difference is that we have the option to do these interim analyses whenever we want and however many times we want. The model is not bound by strict rules which makes it more flexible. If there has to be a decision in the, given the performance of Gombay's **Test 2** and its almost unbounded flexibility, the author sides with a fully sequential approach.

4.5 Future Work

The results of this thesis are based on the comparison of the research from Gombay (2001)and the group sequential methods proposed in Jennison and Turnbull (1999). I made the assumption of exponential survival times for comparisons sake and simplicity. I am interested and have begun researching the results and properties of other distributions such as Weibull. As well it would be interesting to see the results of tests using the canonical joint distribution theory for other distributions and other test statistics.

Chapter 5

Appendix

5.1 Tables

1. Table 5.1: Maximum sample size multipliers for P
2. Table 5.2: Maximum sample size multipliers for OBF
3. Table 5.3: P constants for two-sided tests with overall significance level α
4. Table 5.4: OBF constants for two-sided tests with overall significance level α
5. Table 5.5: Critical Values for the Pure Sequential Tests
6. Table 5.6: Values of $Z_{1-\alpha/2}$ used to calculate M
7. Table 5.7: Certain Values of M

Table 5.1: *Maximum sample size multipliers for Pocock.*

$R_P(K, \alpha, \beta)$						
	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
	$\alpha =$			$\alpha =$		
K	0.01	0.05	0.1	0.01	0.05	0.1
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.092	1.11	1.121	1.084	1.100	1.110
3	1.137	1.166	1.184	1.125	1.151	1.166
4	1.166	1.202	1.224	1.152	1.183	1.202
5	1.187	1.229	1.254	1.170	1.207	1.228
10	1.301	1.301	1.337	1.222	1.271	1.302
20	1.291	1.363	1.411	1.264	1.327	1.367

Table 5.2: *Maximum sample size multipliers for OBF.*

$R_{OBF}(K, \alpha, \beta)$						
	$1 - \beta = 0.8$			$1 - \beta = 0.9$		
	$\alpha =$			$\alpha =$		
K	0.01	0.05	0.1	0.01	0.05	0.1
1	1.000	1.000	1.000	1.000	1.000	1.000
2	1.001	1.008	1.016	1.001	1.007	1.014
3	1.007	1.017	1.027	1.006	1.016	1.025
4	1.011	1.024	1.035	1.010	1.022	1.032
5	1.015	1.028	1.040	1.014	1.026	1.037
10	1.024	1.040	1.053	1.022	1.037	1.049
20	1.030	1.047	1.061	1.029	1.045	1.057

Table 5.3: *Pocock constants for two-sided tests with overall significance level α..*

$C_P(K, \alpha)$			
K	0.01	0.05	0.1
1	2.576	1.96	1.645
2	2.772	2.178	1.875
3	2.873	2.289	1.992
4	2.939	2.361	2.067
5	3.986	2.413	2.122
10	3.117	2.555	2.27
20	3.225	2.672	2.392

Table 5.4: *O'Brien and Flemming constants for two-sided tests with overall significance level α.*

$C_{OBF}(K, \alpha)$			
K	0.01	0.05	0.1
1	2.576	1.960	1.645
2	2.580	1.977	1.678
3	2.595	2.004	1.710
4	2.609	2.024	1.733
5	2.621	2.040	1.751
10	2.660	2.087	1.801
20	2.695	2.126	1.842

Table 5.5: *Critical Values for the Pure Sequential Tests*

	Test 1: $CV_1(n, \alpha)$			Test 2: $CV_2(n, \alpha)$		
n	0.01	0.05	0.1	0.01	0.05	0.1
100	4.570	3.637	3.226	2.80	2.24	1.96
150	4.558	3.650	3.249	2.80	2.24	1.96
200	4.551	3.659	3.265	2.80	2.24	1.96
250	4.547	3.666	3.276	2.80	2.24	1.96

Table 5.6: *Values of $Z_{1-\alpha/2}$ used to calculate M*

	$\alpha =$		
$1 - \beta$	0.01	0.05	0.1
0.5	2.576	1.96	1.645
0.75	3.2503	2.6345	2.3194
0.8	3.4174	2.8016	2.4865
0.9	3.8574	3.2416	2.9264
0.95	4.2207	3.6049	3.2898

Table 5.7: *Certain Values of M*

	$1-\beta = 0.80$			$1-\beta = 0.90$		
	$\alpha =$			$\alpha =$		
K	0.01	0.05	0.10	0.01	0.05	0.10
1	3.417	2.802	2.487	3.857	3.242	2.926
2	2.416	1.981	1.759	2.727	2.292	2.069
3	1.973	1.618	1.436	2.227	1.872	1.689
4	1.709	1.401	1.244	1.929	1.621	1.463
5	1.528	1.253	1.112	1.724	1.450	1.309
10	1.081	0.886	0.786	1.220	1.025	0.925
20	0.764	0.627	0.556	0.862	0.725	0.654

5.2 Comparison 1 Results

1. Table 5.8 Power and Average Stopping Time for Comparison 1, $n = 100$
2. Table 5.9 Power and Average Stopping Time for Comparison 1, $n = 150$
3. Table 5.10 Power and Average Stopping Time for Comparison 1, $n = 200$
4. Table 5.11 Power and Average Stopping Time for Comparison 1, $n = 250$
5. Graph 5.1 Stopping Boundaries for $K = 5$
6. Graph 5.2 Stopping Boundaries for $K = 10$

Table 5.8: *Power and average stopping time for n=100. $\alpha = 0.05$.*

	K	2		5		10	
λ_h		Stop	Power	Stop	Power	Stop	Power
1.0	Test2	99.01	0.0485	99.01	0.0485	99.01	0.0485
	Pocock	98.64	0.0498	97.71	0.0491	97.47	0.0498
	OBF	99.76	0.0519	99.31	0.0509	99.14	0.0514
1.5	Test2	94.23	0.3206	94.23	0.3206	94.23	0.3206
	Pocock	95.17	0.3258	93.51	0.2645	93.26	0.2292
	OBF	98.75	0.3824	95.79	0.3666	94.83	0.3489
2.0	Test2	84.45	0.7420	84.45	0.7420	84.45	0.7420
	Pocock	89.35	0.7463	84.64	0.6806	83.83	0.6229
	OBF	96.43	0.8063	88.45	0.7857	85.90	0.7837

Results for Comparison 1 using U-statistics. Average stopping time is measured in terms of the number of patients to exit the study, either by censoring or death.

Table 5.9: *Power and average stopping time for n=150. $\alpha = 0.05$.*

	K	2		5		10	
λ_h		Stop	Power	Stop	Power	Stop	Power
1.0	Test2	148.52	0.0483	148.52	0.0483	148.52	0.0483
	Pocock	147.70	0.0515	146.42	0.0499	146.06	0.0501
	OBF	149.68	0.0517	148.95	0.0502	148.71	0.0518
1.5	Test2	136.62	0.4630	136.62	0.4630	136.62	0.4630
	Pocock	140.19	0.4667	135.79	0.3934	134.80	0.3575
	OBF	146.97	0.5330	140.44	0.5137	138.00	0.5068
2.0	Test2	114.69	0.9010	114.69	0.9010	114.69	0.9010
	Pocock	125.49	0.9072	115.57	0.8657	112.97	0.8352
	OBF	139.58	0.9372	123.19	0.9286	118.14	0.9218

Results for Comparison 1 using U-statistics. Average stopping time is measured in terms of the number of patients to exit the study, either by censoring or death.

Table 5.10: *Power and average stopping time for n=200. $\alpha = 0.05$.*

	K	2		5		10	
λ_h		Stop	Power	Stop	Power	Stop	Power
1.0	Test2	198.10	0.0498	198.01	0.0498	198.10	0.0498
	Pocock	197.03	0.0503	195.16	0.0499	194.55	0.0510
	OBF	199.49	0.0506	198.59	0.0508	198.20	0.0519
1.5	Test2	175.06	0.5957	175.06	0.5957	175.06	0.5957
	Pocock	182.43	0.5886	174.10	0.5255	173.33	0.4806
	OBF	194.06	0.6590	181.71	0.6450	178.11	0.6331
2.0	Test2	140.18	0.9668	140.18	0.9668	140.18	0.9668
	Pocock	156.91	0.9702	140.22	0.9492	136.56	0.9331
	OBF	180.19	0.9825	153.15	0.9768	145.67	0.9758

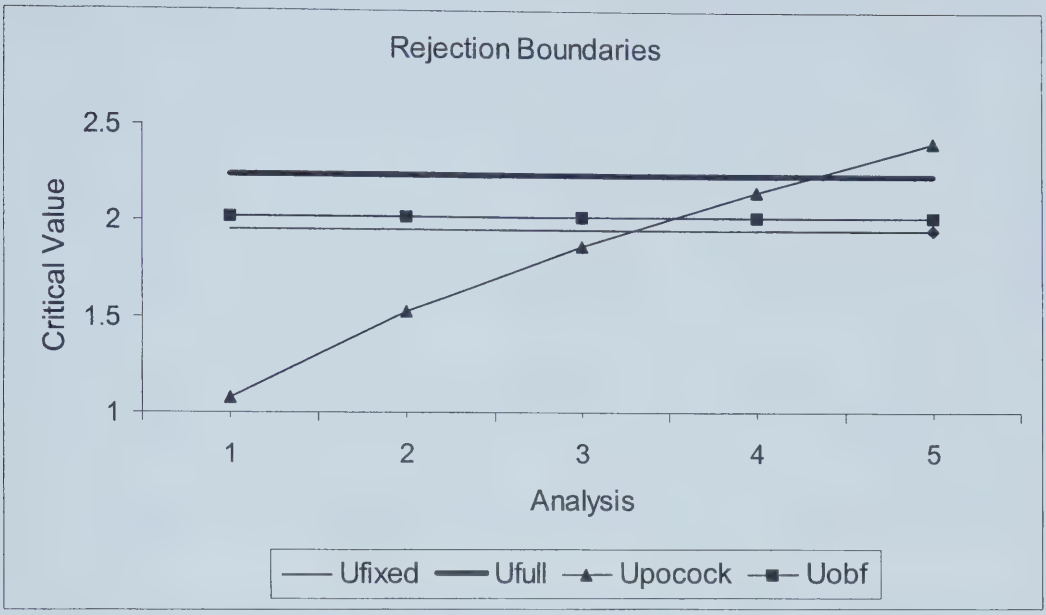
Results for Comparison 1 using U-statistics. Average stopping time is measured in terms of the number of patients to exit the study, either by censoring or death.

Table 5.11: *Power and average stopping time for n=250. $\alpha = 0.05$.*

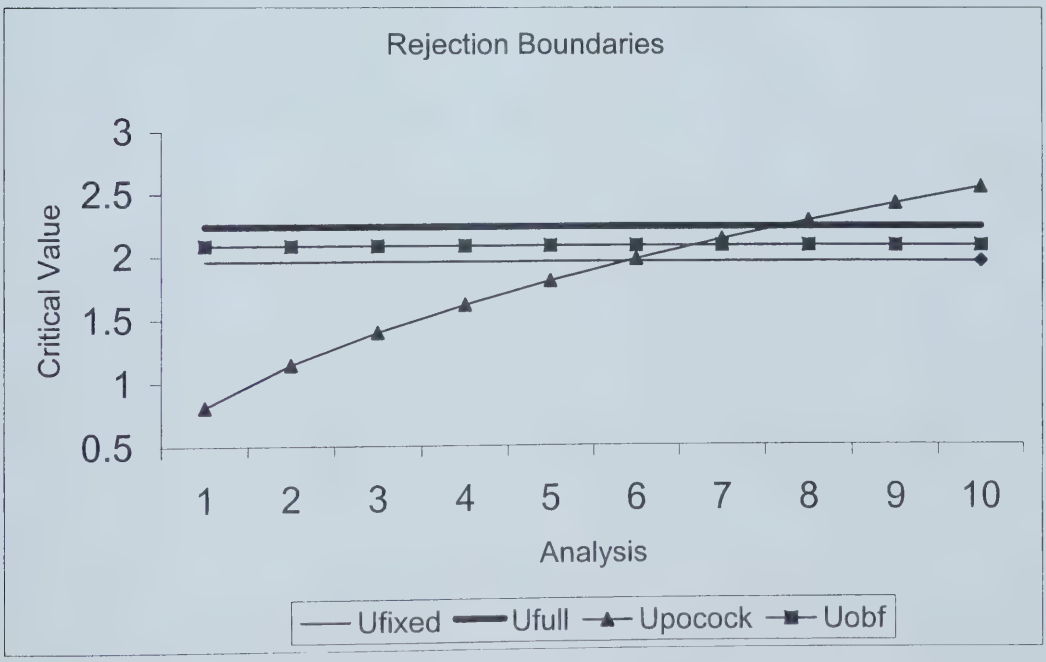
	K	2		5		10	
λ_h		Stop	Power	Stop	Power	Stop	Power
1.0	Test2	247.51	0.0492	247.51	0.0492	247.51	0.0492
	Pocock	246.25	0.0503	243.82	0.0499	243.16	0.0502
	OBF	249.35	0.0512	248.17	0.0509	247.92	0.0514
1.5	Test2	210.94	0.6945	210.94	0.6945	210.94	0.6945
	Pocock	222.95	0.6925	210.81	0.6242	207.89	0.5812
	OBF	240.95	0.7536	221.72	0.7371	214.68	0.7347
2.0	Test2	161.76	0.9918	161.76	0.9918	161.76	0.9918
	Pocock	183.51	0.9921	161.20	0.9836	154.11	0.9804
	OBF	213.91	0.9953	179.23	0.9943	169.26	0.9914

Results for Comparison 1 using U-statistics. Average stopping time is measured in terms of the number of patients to exit the study, either by censoring or death.

Graph 5.1 Stopping Boundaries for Test 2, K=5



Graph 5.2 Stopping Boundaries for Test 2, K=10



5.3 Comparison 2 Results

1. Table 5.12: Power of Test 2 where P has 1 - $\beta = 0.9$
2. Table 5.13: Power of Test 2 where P has 1 - $\beta = 0.8$
3. Table 5.14: Power of Test 2 where OBF has 1 - $\beta = 0.9$
4. Table 5.15: Power of Test 2 where OBF has 1 - $\beta = 0.8$
5. Table 5.16: Power and Average Sample Size at $\alpha = 0.05$, $n = 100$, $K = 5$
6. Table 5.17: Power and Average Sample Size at $\alpha = 0.05$, $n = 150$, $K = 5$
7. Table 5.18: Power and Average Sample Size at $\alpha = 0.05$, $n = 200$, $K = 5$.
8. Table 5.19: Power and Average Sample Size at $\alpha = 0.05$, $n = 250$, $K = 5$
9. Table 5.20: Expected deaths to the observed deaths

Table 5.12: *Power of Test 2 where Pocock has power = 0.9*

	K	2		5		10	
λ_h	λ_A	n	Power	n	Power	n	Power
1.75	1.0	181	0.818	198	0.836	209	0.854
	1.5	200	0.833	219	0.840	231	0.866
	2.0	218	0.845	239	0.854	252	0.880
2.00	1.0	120	0.789	132	0.814	139	0.838
	1.5	130	0.790	143	0.832	151	0.847
	2.0	142	0.819	156	0.843	164	0.858

Every combination of $\{\lambda_h, \lambda_A, K\}$ at a truncation point n (the maximum number of allowed patients) with α=0.05 will achieve a power of 0.9 with Pocock boundaries for the group sequential log-rank test (the best case scenario). The table illustrates the power achieved by ***Test 2*** *using U-statistics.*

Table 5.13: *Power of Test 2 where Pocock has power = 0.8*

	K	2		5		10	
λ_h	λ_A	n	Power	n	Power	n	Power
1.75	1.0	136	0.657	151	0.707	160	0.756
	1.5	151	0.676	167	0.732	177	0.761
	2.0	163	0.685	182	0.743	193	0.779
2.00	1.0	91	0.657	100	0.692	106	0.715
	1.5	98	0.669	109	0.708	115	0.760
	2.0	107	0.677	119	0.725	126	0.792

Every combination of $\{\lambda_h, \lambda_A, K\}$ at a truncation point n (the maximum number of allowed patients) with α=0.05 will achieve a power of 0.8 with Pocock boundaries for the group sequential log-rank test (the best case scenario). The table illustrates the power achieved by ***Test 2*** *using U-statistics.*

Table 5.14: *Power of Test 2 where OBF has power = 0.9*

	K	2		5		10	
λ_h	λ_A	n	Power	n	Power	n	Power
1.75	1.0	166	0.744	169	0.788	170	0.771
	1.5	183	0.789	186	0.790	188	0.771
	2.0	200	0.801	203	0.812	206	0.796
2.00	1.0	110	0.758	112	0.757	113	0.768
	1.5	119	0.786	122	0.775	123	0.772
	2.0	130	0.754	133	0.773	134	0.779

Every combination of $\{\lambda_h, \lambda_A, K\}$ at a truncation point n (the maximum number of allowed patients) with α=0.05 will achieve a power of 0.9 with OBF boundaries for the group sequential log-rank test (the best case scenario). The table illustrates the power achieved by ***Test 2*** *using U-statistics.*

Table 5.15: Power of Test 2 where OBF has power = 0.8

	K	2		5		10	
λ_h	λ_A	n	Power	n	Power	n	Power
1.75	1.0	124	0.624	126	0.649	127	0.639
	1.5	137	0.663	140	0.660	141	0.641
	2.0	149	0.671	152	0.690	153	0.661
2.00	1.0	82	0.621	84	0.628	85	0.626
	1.5	89	0.629	91	0.652	92	0.638
	2.0	98	0.655	99	0.673	101	0.654

Every combination of $\{\lambda_h, \lambda_A, K\}$ at a truncation point n (the maximum number of allowed patients) with α=0.05 will achieve a power of 0.8 with OBF boundaries for the group sequential log-rank test (the best case scenario). The table illustrates the power achieved by ***Test 2*** *using U-statistics.*

Table 5.16: *Power and Average Number of Deaths at $\alpha = 0.05$, n=100, K=5.*

		Expected # of Deaths	Test 2		Sequential Log-Rank: Pocock		Sequential Log-Rank: OBF	
λ_h	λ_A		Power	Deaths	Power	Deaths	Power	Deaths
1.0	1.0	85.5	0.475	84.97	0.05	83.87	0.05	85.39
1.5	1.0	83.22	0.3068	80.24	0.362	71.83	0.443	77.00
	1.5	77.33	0.2827	75.99	0.339	67.92	0.423	71.15
	2.0	71.82	0.2701	70.94	0.315	63.33	0.390	66.91
2.0	1.0	80.85	0.6985	73.90	0.799	52.47	0.864	61.14
	1.5	74.22	0.6676	70.56	0.765	50.23	0.838	57.89
	2.0	68.12	0.6387	66.38	0.727	47.79	0.805	54.48
2.5	1.0	78.51	0.9158	67.56	0.960	38.28	0.979	49.32
	1.5	71.32	0.8742	66.77	0.942	36.94	0.968	46.63
	2.0	65.03	0.8648	63.11	0.936	35.64	0.964	44.19

The first column is the expected number of deaths in a fixed sample setting, with a truncation point n=100. The power and average number of uncensored deaths is simulated for Test 2 using U-statistics against the best case scenario of the group sequential log-rank statistics using Pocock, and O'Brien and Fleming boundaries.

Table 5.17: *Power and Average Number of Deaths at $\alpha = 0.05$, n=150, K=5.*

		Expected # of Deaths	Test 2		Pocock		OBF	
λ_h	λ_A		Power	Deaths	Power	Deaths	Power	Deaths
1.0	1.0	128.25	0.0470	127.49	0.05	125.32	0.05	127.58
1.5	1.0	124.83	0.4432	118.38	0.513	100.03	0.604	108.99
	1.5	116.00	0.4262	112.87	0.482	94.47	0.572	102.30
	2.0	107.73	0.3951	105.57	0.453	89.24	0.541	96.16
2.0	1.0	121.28	0.8728	103.66	0.937	63.75	0.965	79.95
	1.5	111.34	0.8494	102.51	0.915	61.51	0.950	75.81
	2.0	102.19	0.8286	98.23	0.889	59.25	0.933	71.86
2.5	1.0	117.76	0.9836	92.74	0.996	44.11	0.998	62.43
	1.5	106.98	0.9715	94.90	0.992	42.63	0.997	58.96
	2.0	97.54	0.9690	92.93	0.986	41.36	0.994	56.02

The first column is the expected number of deaths in a fixed sample setting, with a truncation point n=150. The power and average number of uncensored deaths is simulated for Test 2 using U-statistics against the best case scenario of the group sequential log-rank statistics using Pocock, and O'Brien and Fleming boundaries.

Table 5.18: *Power and Average Number of Deaths at $\alpha = 0.05$, n=200, K=5.*

λ_h	λ_A	Expected # of Deaths	Test 2		Pocock		OBF	
			Power	Deaths	Power	Deaths	Power	Deaths
1.0	1.0	171.00	0.0472	169.13	0.05	166.77	0.05	169.78
	1.0	166.44	0.5682	154.02	0.664	123.61	0.731	137.95
1.5	1.5	154.66	0.5481	147.80	0.609	117.38	0.698	130.05
	2.0	143.64	0.5273	140.34	0.576	111.34	0.666	122.55
	1.0	161.71	0.9548	128.97	0.982	70.88	0.990	94.75
2.0	1.5	148.45	0.9401	131.38	0.973	68.80	0.987	90.20
	2.0	136.25	0.9258	129.25	0.960	66.71	0.979	85.79
	1.0	157.02	0.9964	112.89	0.9997	49.15	0.9999	74.27
2.5	1.5	142.64	0.9950	119.96	0.9992	47.29	0.9997	69.93
	2.0	130.06	0.9949	119.64	0.9982	45.78	0.9993	66.39

The first column is the expected number of deaths in a fixed sample setting, with a truncation point n=200. The power and average number of uncensored deaths is simulated for Test 2 using U-statistics against the best case scenario of the group sequential log-rank statistics using Pocock, and O'Brien and Fleming boundaries.

Table 5.19: *Power and Average Number of Deaths at $\alpha = 0.05$, n=250, K=5.*

λ_h	λ_A	Expected # of Deaths	Test 2		Pocock		OBF	
			Power	Deaths	Power	Deaths	Power	Deaths
1.0	1.0	213.75	0.0487	212.09	0.05	208.46	0.05	212.22
	1.0	208.50	0.6695	189.26	0.748	143.14	0.822	163.46
1.5	1.5	193.33	0.6503	181.99	0.714	136.36	0.793	154.71
	2.0	179.55	0.6053	174.37	0.679	129.96	0.762	146.18
	1.0	202.14	0.9834	151.95	0.996	76.66	0.998	108.11
2.0	1.5	185.56	0.9773	159.06	0.992	74.36	0.997	102.72
	2.0	170.31	0.9689	156.41	0.987	72.23	0.994	97.80
	1.0	196.27	0.9996	130.98	0.9999	54.02	1.0000	85.25
2.5	1.5	178.30	0.9994	139.87	0.9999	51.76	1.0000	80.22
	2.0	162.57	0.9991	144.26	0.9998	49.77	0.9999	75.70

The first column is the expected number of deaths in a fixed sample setting, with a truncation point n=250. The power and average number of uncensored deaths is simulated for Test 2 using U-statistics against the best case scenario of the group sequential log-rank statistics using Pocock, and O'Brien and Fleming boundaries.

Table 5.20: A comparison of the expected number of deaths to the observed number of deaths after censoring.

n	Lambda	LambdaC	LambdaT	Expected Deaths	Observed Deaths
100	1.5	1	1.5	83.22	83.29
100	1.5	1.5	2.25	77.33	77.36
100	1.5	2	3	71.82	71.87
100	2	1	2	80.85	80.92
100	2	1.5	3	74.22	74.19
100	2	2	4	68.12	67.94
100	2.5	1	2.5	78.51	78.37
100	2.5	1.5	3.75	71.32	71.16
100	2.5	2	5	65.03	65.15
150	1.5	1	1.5	124.83	124.87
150	1.5	1.5	2.25	116.00	116.60
150	1.5	2	3	107.73	107.58
150	2	1	2	121.28	121.18
150	2	1.5	3	111.34	111.26
150	2	2	4	102.19	102.41
150	2.5	1	2.5	117.76	117.88
150	2.5	1.5	3.75	106.98	107.16
150	2.5	2	5	97.54	97.69
200	1.5	1	1.5	166.44	166.10
200	1.5	1.5	2.25	154.66	154.37
200	1.5	2	3	143.64	143.75
200	2	1	2	161.71	161.44
200	2	1.5	3	148.45	148.32
200	2	2	4	136.25	136.26
200	2.5	1	2.5	157.02	157.36
200	2.5	1.5	3.75	142.64	142.76
200	2.5	2	5	130.06	130.00
250	1.5	1	1.5	208.05	208.02
250	1.5	1.5	2.25	193.33	193.35
250	1.5	2	3	179.55	179.50
250	2	1	2	202.14	202.36
250	2	1.5	3	185.56	185.62
250	2	2	4	170.31	170.47
250	2.5	1	2.5	196.27	196.67
250	2.5	1.5	3.75	178.30	178.06
250	2.5	2	5	162.57	162.73

5.4 Glossary

α - Significance level

α_i' - Nominal significance levels in the group sequential setting, $i = 1, \ldots, K$

$1 - \beta$ - Power of the test

γ - General distribution parameter used in Weibull description

Γ - Gamma function

δ - Value of θ under the alternative hypothesis

Δ - Indicator variable to indicate censored observation

θ - Some population parameter

κ - Boundary index in Wang and Tsiatis (Section 3.2.1)

λ - Distribution parameter used in Exponential and Weibull description

λ_A - Hazard rate of an exponential response from treatment A (sometimes referred to as the control group)

λ_B - Hazard rate of an exponential response from treatment B (sometimes referred to as the treatment group)

λ_h - Hazard ratio $= \lambda_A/\lambda_B$

μ - Population mean from a normal distribution

π - Indicator variable to indicate treatment allocation

ρ - Used in section 3.1.4 (Weibull)

σ - Population standard deviation for a normal distribution

$\sum$ - The sum of

τ - Incremental information in an equal group setting for group sequential analysis, (Section 3.2.1)

φ - Probability distribution for treatment allocation, $P(\pi = 1) = \varphi$

ω_1 - Probability of type I censoring

ω_2 - Probability of type II censoring

K - The number of interim analysis in the group sequential setting

k' - The time of analysis in the group sequential setting, $k' = 1, \ldots, K$

k - The time of analysis in a fully sequential setting (measured by the number of accrued events), $k = 1, \ldots, n$, $k = k_A + k_B$. In a group sequential setting $k = (n/K)k'$

k_A - The number of accrued events in group A at time k

k_B - The number of accrued events in group B at time k

n - Truncation point. This is the maximum number of patients allowed in the study. $n = n_A + n_B$

n_A - The number of observations in group A

n_B -The number of observations in group B

N - The total number of uncensored events. $N = N_A + N_B$.

N_A - The total number of uncensored events in group A

N_B - The total number of uncensored events in group B

$N_{k'}$ - The number of uncensored events (deaths) at analysis k'

N_g - The number of uncensored events per group per analysis in an equal groups model for group sequential testing

C_P - Critical values for Pocock. (Table 5.3) P = Pocock

C_{OBF} - Critical values for O'Brien and Fleming. (Table 5.4) OBF = O'Brien and Fleming

CV_1 and CV_2 - Critical values for **Test 1** and **Test 2** respectively. (Table 5.5)

M - A multiplier defined in formula (3.8). (Table 5.7)

R - A multiplier to determine maximum sample size for group sequential models. (Tables 5.1, 5.2)

Z_α - The cumulative distribution function of a standard normal distribution at α.

$N(0,1)$ - The standard normal distribution

$U_j^{(y)}$ - The entry time of patient j in treatment y.

$T_j^{(y)}$ - The exact survival time of patient j in treatment y.

$C_j^{(y)}$ - The censored survival time of patient j in treatment y.

$X_j^{(y)}$ - The recorded survival time of patient j in treatment y, $\min(X, C)$

$E_j^{(y)}$ - The event time of patient j in treatment y.

5.5 Code

```
#include <stdlib.h>
#include <stdio.h>
#include <time.h>
#include <conio.h>
#include <math.h>

FILE * matrixFile;
const NUM = User Defined;

const float lambda1 = User Defined;
const float lambda0 = User Defined;

const look = User Defined;

int i,j,k,n,l, t1, t2, lost, maxdeaths;

int ObsTest2, ObsPoc, ObsOBF, bTest2, bPoc, bOBF;
struct RanMatrix{
float randentry;
float entry;
float randsurvival;
double survival;
float event;
float randcensor;
int censor;
float randtreat;
int treat;
};

void RanMatrixSort(struct RanMatrix vec[NUM]);
void RanMatrixPrint(struct RanMatrix vec[NUM]);

double RandDbl(void)
{
return rand() / ((double)RAND_MAX+1.0);
}
```

```
main()
{
matrixFile = fopen("C:\Matrix.txt",
                                                    "w");
if (!matrixFile)
        {puts("An error occured while opening the file!");
        exit(1);
  }

struct RanMatrix vec[NUM];

time_t t;
srand((unsigned) time(&t));

for (j=0; j<10000; j++)    {
printf("%d", j);
printf("\n");

vec[i].randentry=RandDbl();
vec[i].entry=10.0*vec[i].randentry;

vec[i].randtreat=RandDbl();
if (vec[i].randtreat <= 0.5) {vec[i].treat=1;}
else {vec[i].treat = 0;}

/*Survival times are Exp(lambda1), Exp(lambda0) for trt 1 and 0 resp. */

vec[i].randsurvival=RandDbl();
if (vec[i].randsurvival>0 && vec[i].treat==1)
{vec[i].survival=-lambda1*log(vec[i].randsurvival);}
else
if (vec[i].randsurvival>0 && vec[i].treat==0)
{vec[i].survival=-lambda0*log(vec[i].randsurvival);}

vec[i].event=vec[i].entry+vec[i].survival;

vec[i].randcensor=RandDbl();
if (vec[i].randcensor<=0.05)
{vec[i].censor=1;}
else {vec[i].censor=0;}

if (vec[i].randcensor<=0.05) lost = lost +1;
maxdeaths = NUM*100000 - lost;

}
```

```
const float alpha = 0.05;
const float lambda = 0.5;

float POC, CV2, OBF;

int Hmatrix[NUM][NUM];
int Umatrix[NUM][NUM];
int UmatrixAll[NUM][NUM];

float SSstat[NUM];
float Ustat[NUM];
float TestStat1[NUM];


/* CV for look=5 */
CV2 = 2.24;
OBF = 2.087;
POC = 2.555;

float CVOBF[look];
float CVPOC[look];

for (i=1; i<look+2; i++)
{CVPOC[i] = 1.0* POC * pow(i*0.1, 0.5);}



   /* Create the Hmatrix. */
   for (k=0; k<NUM; k++)
         {
     for (i=0; i<NUM; i++)
     if ((vec[i].survival < vec[k].survival) && (vec[i].censor==0)) Hmatrix[i][k]=1;
     else
     if ((vec[i].survival == vec[k].survival) && (vec[i].censor==0) &&
(vec[k].censor==1))                Hmatrix[i][k]=1;
     else
     if ((vec[i].survival > vec[k].survival) && (vec[k].censor==0)) Hmatrix[i][k]=-1;
     else
     if ((vec[i].survival == vec[k].survival) && (vec[k].censor==0) &&
(vec[i].censor==1)) Hmatrix[i][k]=-1;
     else
     if ((vec[i].survival == vec[k].survival) && (vec[i].censor==1)) Hmatrix[i][k]=0;
     else
     Hmatrix[i][k]=0;
     }
```

```
/* Calculate the Umatrix for all treatments by summing the appropriate values in
the Hmatrix. */
for (i=0; i<NUM; i++) {
                    Umatrix[-1][i]=0;}
for (k=0; k<NUM; k++) {
for (i=0; i<NUM; i++) {
 if ((vec[k].treat==1)) Umatrix[i][k] = Hmatrix[i][k] + Umatrix[i-1][k];
 else Umatrix[i][k]=0;}
 }

/* Calculate the Umatrix for treatment 1 by summing the appropriate values in
the Hmatrix. */
for (i=0; i<NUM; i++) {
   UmatrixAll[-1][i]=0;}
for (k=0; k<NUM; k++) {
for (i=0; i<NUM; i++) {
 UmatrixAll[i][k] = Hmatrix[i][k] + UmatrixAll[i-1][k];}
 }


/*reset Ustat */
for (i=0; i<NUM; i++)
 {Ustat[i] = 0;}

/* Calculate Ustatistic. */
for (i=0; i<NUM; i++) {
    for (k=0; k<NUM; k++)
                    if (k<=i) Ustat[i]=Ustat[i]+Umatrix[i][k];
  }

/* Reset SS */
for (i=0; i<NUM; i++)
    {SSstat[i] = 0;}

/*Calculate Sum of squares for all i. */
for (i=0; i<NUM; i++) {
    for (k=0; k<NUM; k++)
  if (k<=i) SSstat[i] = SSstat[i] + (UmatrixAll[i][k]*UmatrixAll[i][k]);
  }

/*Calculate Test Stat */
  for (i=0; i < NUM; i++)  {
if (Ustat[i]<=0 && SSstat[i]>0) TestStat1[i] = (-Ustat[i])*(pow(SSstat[i], -
0.5))*(pow(0.5, -1.0))*(pow(i+1,0.5))*(pow(NUM,-0.5));
else
```

```
    if (Ustat[i]>0 && SSstat[i]>0) TestStat1[i] = (Ustat[i])*(pow(SSstat[i], -
0.5))*(pow(0.5, -1.0))*(pow(i+1,0.5))*(pow(NUM,-0.5))
    ;
    else TestStat1[i] = 0;
    }

    for (l = 1; l <look+1 ; l++) {
        t1 = l*(NUM/look) ;
    if (TestStat1[t1-1] > OBF) goto Output;}

  Output:
        for (i=0; i<NUM; i++) {
    if (TestStat1[i] > CV2) goto Output2;}

  Output2:
    for (n=1; n<look+1; n++) {
      t2 = n*(NUM/look);
    if (TestStat1[t2-1] > CVPOC[n]) goto Output3;}

  Output3:

 /* if (i<NUM-2) bTest2 = bTest2+1;
  else
  if (i=NUM-1 && TestStat1[NUM-1] > CV2) bTest2 = bTest2 +1;

  if (t1<NUM) bOBF = bOBF +1;
  else
  if (t1=NUM && TestStat1[NUM-1] > OBF) bOBF = bOBF +1;

  if (t2<NUM) bPoc - bPoc + 1;
  else
  if (t2=NUM && TestStat1[NUM-1] > CVPOC[look]) bPoc = bPoc + 1;
   */

   if (TestStat1[i]>CV2) bTest2 = bTest2+1;
   if (TestStat1[t1-1] > OBF) bOBF = bOBF +1;
   if (TestStat1[t2-1] > CVPOC[n]) bPoc = bPoc +1;


  ObsTest2 = ObsTest2 + i+1;
  ObsPoc = ObsPoc + t2;
  ObsOBF = ObsOBF + t1;
```

```
   if (TestStat1[t1-1] > OBF) fprintf(matrixFile, "OBF: The test is rejected at n=
%2d", t1);
   else
   if (TestStat1[NUM-1] > OBF) fprintf(matrixFile, "OBF: The test is rejected at n=
%2d", NUM);
   else fprintf(matrixFile, "OBF: The test was not rejected.");

   fprintf(matrixFile, "\n");

   if (TestStat1[i] > CV2) fprintf(matrixFile, "Test2: The test is rejected at n= %2d",
i+1);
   else fprintf(matrixFile, "Test2: The test was not rejected.");


   fprintf(matrixFile, "\n");

   if (TestStat1[t2-1] > CVPOC[n]) fprintf(matrixFile, "POC: The test is rejected at
n= %2d", t2);

   else fprintf(matrixFile, "POC: The test was not rejected.");
              */
  /*fprintf(matrixFile, "\n");  */
            /*
 fprintf(matrixFile, "\n");
   fprintf(matrixFile, "%10d %10d %10f %10f", t2, n, TestStat1[t2-1], CVPOC[n]);
              */

 }


 ObsTest2 = ObsTest2 - (100000-bTest2);
fprintf(matrixFile, "\n");
fprintf(matrixFile, "\n");
fprintf(matrixFile, "%10d %10d %10d", ObsTest2, ObsPoc, ObsOBF);
fprintf(matrixFile, "\n");
fprintf(matrixFile, "%10d %10d %10d", bTest2, bPoc, bOBF);

fprintf(matrixFile, "\n");
fprintf(matrixFile, "%10d %10d", lost, maxdeaths);

fclose(matrixFile);

getch();
return 0 ;
}
```

```
//*************************************************//
void RanMatrixSort(struct RanMatrix vec[NUM])
{
int inner, outer;
struct RanMatrix temp;
int didSwap;
for (outer=0; outer <(NUM-1); outer++)
        {
  didSwap=0;
  for (inner=outer; inner<NUM; inner++)
        {
    if (vec[outer].event > vec[inner].event)
        {temp = vec[outer];
       vec[outer] = vec[inner];
       vec[inner] = temp;
       didSwap = 1;
       }
    }
  if (!didSwap)
        {break;}
  }
return;
}

//*************************************************//
void RanMatrixPrint(struct RanMatrix vec[NUM])
{
        int ctr = 0;
  for (ctr = 0; ctr < NUM; ctr++)
                {printf("%10f %14f %7d %7d %14f\n",
    vec[ctr].entry, vec[ctr].survival, vec[ctr].treat,
    vec[ctr].censor, vec[ctr].event);
    }
  return;
}
```

Chapter 6

Bibliography

1. Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika, 52, 203 - 223.

2. Gombay, E. and Heo, G. (2001). Nonparametric Tests for Comparing Two Treatments in Sequential Trials. Technical Paper, University of Alberta.

3. Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution. Annals of Mathematics and Statistics, 19, 293 - 325.

4. Jennison and Turnbull (1999). Group Sequential Methods with Application to Clinical Trials. Chapman Hall.

5. Mantel, N. (1967). Ranking Procedures for Arbitrarily Restricted Observations. Biometrics, 23, 65 - 78.

6. Miller R. G. (1981). Survival Analysis. Wiley Classics Library.

7. O'Brien P. C. and Fleming T. R. (1979). A Multiple Testing Procedure for Clinical Trials. Biometrics, 35, 549 - 556.

8. Pocock, S. J. (1977). Group Sequential methods in the design and analysis of clinical trials. Biometrika, 64, 191 - 199.

9. S+SeqTrial (2001). Software for Design and Analysis of Sequential Trials. Insightful Corp.

10. Wang, S.K. and Tsiatis, A.A. (1987). Approximately optimal one-paramter boundaries for group sequential trials. Biometrics, 43, 193-200.

University of Alberta Library
0 1620 1748 6315
B45837